INTRINSIC DIMENSION IN FROZEN WEIGHT SPACE: DISTORTION AT SOLUTION, COLLAPSE AT ORIGIN

Anonymous authors
Paper under double-blind review

ABSTRACT

A neural network can be trained through a few coordinates mapped into its weights by a fixed low-dimensional map, a *chart*. The smallest such chart that solves a task is reported as the task's intrinsic dimension, and the same construction now generates weights outright. A basic question has gone unasked: is a chart that is well behaved in weight space enough to give a well-behaved optimization problem in the trained coordinates? We study it where the chart is fixed, random and known exactly, and find that it is not. The architecture reshapes the mapped directions in two ways. It *distorts* them: at trained solutions a handful of directions carry almost all of the pullback Fisher mass, and a better-oriented chart helps only slightly, at intermediate budgets. And it can *collapse* them: an operation that multiplies two generated matrices, as attention multiplies its query and key, responds to the chart only at higher order near a zero base point, so attention barely moves while the rest of the network trains around it. We characterize collapse by a *functional order*: the power of latent scale in its gradient. The computation graph predicts this integer for charts without leading-term cancellation. Measured slopes follow it on graph and image hosts, approaching it more slowly on a vision transformer; changing which matrices are generated moves the slopes as predicted without adding trainable parameters. Measurements with zero-centred charts, as in current weight generators, therefore depend on the host and the base point as well as the task; the original construction, centred at initialization, avoids collapse but not distortion.

1 INTRODUCTION

A frozen random *chart* maps a unit-norm latent vector across many weights of a *host* network, starting them one to two orders of magnitude below its usual initialization. The obvious repair is to turn the latent vector up. On a graph transformer it does almost nothing: fifty times the latent scale leaves attention within three percent of uniform after training. Leaving the scale alone and changing which matrices are generated moves the same pattern more than twice as far on two of three benchmarks, at no parameter cost. Near the origin, magnitude is not what is wrong.

The issue appears after composing the chart with the host. Intrinsic-dimension measurements freeze a random chart from a small latent space into weight space before seeing data, then train only the latent vector. The smallest chart that solves the task is reported as a task property (Li et al., 2018). That threshold has been used to compare tasks and explain fine-tuning (Aghajanyan et al., 2021); the chart has been reused to train (Gressmann et al., 2020) and generate weights (Sen & Mukherjee, 2026). Li et al. (2018) already report a gap across architectures and read it through minimum description length, the host that needs fewer coordinates being the better matched one. We do not dispute that reading. The same number is equally consistent with a host that has operations the chart cannot move and has trained around them.

A random chart is isotropic in weight space, but the host maps weight directions to function changes: its Jacobian, far from uniform, sets the local geometry, and the products in its computation graph set the order at which each operation responds. Training explores the chart's image under that Jacobian; it can be lower dimensional, distorted, or sent to zero. Random in weights is not random in functions. When an operation depends on two generated matrices only through their product, as an attention score depends on its query and key, each matrix is small near the origin and the product is smaller

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

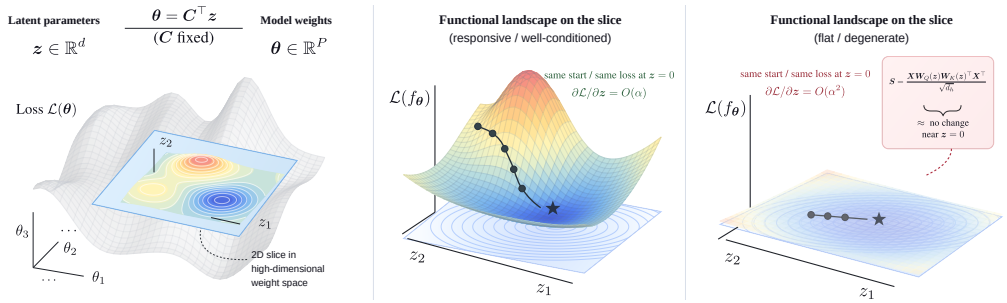


Figure 1: One fixed chart in weight space (left) has different effects after composition with two hosts: responsive (centre), or flat near zero when attention multiplies two generated matrices (right).

still, so the chart does not reach that operation to first order (Figure 1), and the gradient that might carry it away is built from the same vanishing factors. Turning the latent vector up rescales both factors at once, which is why it changes so little.

This paper makes that account exact and tests it. Intrinsic dimension mixes *distortion* (which functional directions a chart represents, and how evenly) with *collapse* (how strongly a generated operation responds near the *base point*, the weights at zero latent). For distortion, treating the chart as a reparameterization of the host separates the directions it supplies in weight space from the ones the host presents to the optimizer, and the pullback Fisher describes the second: at trained solutions on graph and image hosts it is highly anisotropic. For collapse, the exponent at which a generated operation's gradient vanishes near the base point is its *functional order*: at fixed base point and choice of generated tensors, the computation graph predicts it across charts unless their leading terms cancel. On three hosts measured gradient slopes match or approach the predicted integers, and three changes at a fixed budget, freezing one factor, generating a product as a single matrix, and moving the base point, lower the measured order as predicted. Two limits belong here. Collapse needs the base point at zero, where recent weight-generation constructions place it and the original one does not, which we use as a control; for the original construction our claim is limited to distortion. And the order decides which operations a chart can move, not what those operations are worth. We offer none of these changes as a training method. Our contributions are:

- **We show that zero-centred intrinsic dimension mixes task difficulty with architectural accessibility.** The same task and latent budget can induce fundamentally different functional restrictions across hosts.
- **We demonstrate that low intrinsic dimension need not imply an easy task.** A host can train around operations that the chart cannot effectively move, making a low dimension consistent with architectural redundancy or spare capacity.
- **We prove that leading-order accessibility can be predicted before training.** Under non-cancellation, the functional order of each generated operation is set by the computation graph and is independent of the chart; experiments on GT and CNN recover the predicted integer slopes.
- **We demonstrate that parameterization controls accessibility at fixed budget.** Changing what is generated lowers the functional order of the same operation without adding trainable parameters, making weight generation itself a design variable rather than an implementation detail.

2 Related Work

Low-dimensional training subspaces. Random subspaces were introduced to estimate intrinsic dimension (Li et al., 2018), then used for efficient training (Gressmann et al., 2020) and to explain fine-tuning (Aghajanyan et al., 2021). Although Li et al. (2018) relate architectural differences to minimum description length, the estimate also depends on initialization (Fort & Scherlis, 2019), threshold and burn-in (Larsen et al., 2022), perturbation norm (Kim et al., 2026), the subspaces followed by gradient descent (Song et al., 2025), and the orientation of the displacement relative to curvature (Cheng et al., 2026). We add a functional source of variation: composition with the architecture can distort or annihilate an isotropic weight-space chart. Truncatable subspaces explic-

itly preserve prefixes (Rauba & van der Schaar, 2026; Valipour et al., 2023; Modoranu et al., 2026), while adaptive low-rank methods update the subspace (Hu et al., 2022; Zhao et al., 2024); both sacrifice the fixed coordinate system used here to isolate this effect.

Hypernetworks and weight generation. Hypernetworks generate a target model's weights from a smaller network (Ha et al., 2017). Recent work instead uses a frozen orthogonal map, scalar modulation, and weight-space bias (Sen & Mukherjee, 2026). We use this controlled construction to study when generated directions still change predictions, not to propose a new generator.

Information geometry and second-order methods. The Fisher defines the natural metric and natural gradient (Amari, 1998); scalable methods typically approximate it, for example with Kronecker factors (Martens & Grosse, 2015). We apply an unbiased true-Fisher estimator (Sun, 2026) and a local-dimension measure (Sun & Nielsen, 2025) directly to the chart, where the pullback is exact (Kristiadi et al., 2023). Appendix O discusses Fisher spectra and the limits of the empirical Fisher (Karakida et al., 2019; Papyan, 2019; Ghorbani et al., 2019; Kunstner et al., 2019).

Small initialization, homogeneity and factorized optimization. Factorized parameterizations alter optimization geometry (Arora et al., 2018), and small initialization can induce saddle-to-saddle dynamics near low-rank critical manifolds (Jacot et al., 2021). In transformers, query–key factors can move slowly (Vashisht & Ramaswamy, 2026), attention can remain near maximum entropy (Zhai et al., 2023), and its Hessian differs structurally from that of feedforward blocks (Ormaniec et al., 2025). That products of small factors have small gradients is elementary, and these analyses establish slow dynamics for particular factorized models. We instead assign an order to every block of an arbitrary generated computation graph and propagate it through softmax and normalization. Separating that order from chart-span geometry predicts which operations a frozen zero-centred generator cannot move without solving the dynamics.

3 THEORY

One intrinsic-dimension threshold can hide two restrictions. At trained solutions, *distortion* means useful directions are lost or unevenly represented; near a zero base point, *collapse* means an operation's gradient vanishes with latent scale. We measure the first with the Fisher (Sections 3.1, 3.2), predict the second from the computation graph (Section 3.3), then compare their interventions and training effects (Sections 3.4–3.6).

3.1 FROM WEIGHT DIRECTIONS TO FUNCTION CHANGES

Let $\mathcal{L}$ be the objective and $g(\boldsymbol{z}) = \boldsymbol{\theta}_0 + \boldsymbol{C}^\top \boldsymbol{z}$ a frozen chart with orthonormal rows. It maps $\boldsymbol{z} = 0$ to the *base point* $\boldsymbol{\theta}_0$ and is zero-centred only when $\boldsymbol{\theta}_0 = 0$. Its reachable weights are $\boldsymbol{\theta}_0 + \mathcal{S}_C$, where $\mathcal{S}_C = \operatorname{Im} \boldsymbol{C}^\top$, and they decide attainability exactly:

$$\inf_{\boldsymbol{z}} \mathcal{L}(g(\boldsymbol{z})) < \tau \iff (\boldsymbol{\theta}_0 + \mathcal{S}_C) \cap \{\boldsymbol{\theta} : \mathcal{L}(\boldsymbol{\theta}) < \tau\} \neq \varnothing. \tag{1}$$

The trained coordinates affect predictions through $h(\boldsymbol{z}; x) = f_{g(\boldsymbol{z})}(x)$. Their first-order changes span $\mathcal{F}_C(x) = \operatorname{Im}(\boldsymbol{J}_{\boldsymbol{\theta}} f(x) \boldsymbol{C}^\top)$, since $\boldsymbol{J}_{\boldsymbol{z}} h = \boldsymbol{J}_{\boldsymbol{\theta}} f \, \boldsymbol{C}^\top$. Thus d counts weight directions, while the host controls which change predictions. With $\boldsymbol{H}_f$ the output-space Fisher, their sensitivity is measured by the pullback Fisher

$$\boldsymbol{F}_{\boldsymbol{z}} = \boldsymbol{C} \, \boldsymbol{F}_{\boldsymbol{\theta}} \, \boldsymbol{C}^\top = \mathbb{E}_x \big[\boldsymbol{J}_{\boldsymbol{z}} h^\top \boldsymbol{H}_f \, \boldsymbol{J}_{\boldsymbol{z}} h \big]. \tag{2}$$

Write $\boldsymbol{F} = \boldsymbol{F}_{\boldsymbol{\theta}}$. The spectrum of $\boldsymbol{F}_{\boldsymbol{z}}$ measures uneven sensitivity. For local steps, $V_C := \boldsymbol{F}^{1/2} \operatorname{Im} \boldsymbol{C}^\top$ is the Fisher-weighted span the chart reaches; γ below measures retained local decrease. We call these effects *distortion*. *Collapse* instead concerns a block's gradient vanishing along $\boldsymbol{z} = \varepsilon \boldsymbol{u}$.

Definition 1 (Functional order)**.** The exponent of a block's gradient along $\boldsymbol{z} = \varepsilon \boldsymbol{u}$ is $\operatorname{ord}_\ell(\boldsymbol{u}) := \lim_{\varepsilon \to 0} \log \|\nabla_{\boldsymbol{z}_\ell} \mathcal{L}(\varepsilon \boldsymbol{u})\| / \log \varepsilon$, when the limit exists; Theorem 2 shows it is a constant ord_ℓ for almost every $\boldsymbol{u}$. *Collapse* is $\operatorname{ord}_\ell \geq 1$. A latent shared across blocks sums their contributions (Section 3.3).

3.2 GEOMETRY: THE FUNCTIONAL DIRECTIONS A CHART PRESERVES

We ask how much local improvement the chart retains. To compare directions rather than scale, we use the Fisher (Gauss–Newton) local model $\mathcal{Q}(\boldsymbol{\delta}) = \mathcal{L}(\boldsymbol{\theta}_0) + \boldsymbol{g}^\top \boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}^\top \boldsymbol{F}\boldsymbol{\delta}$, not the Taylor model, with $\boldsymbol{F} \succeq 0$ and $\boldsymbol{g} \in \operatorname{Im} \boldsymbol{F}$ (Appendix A.10), $\boldsymbol{g}_z = \boldsymbol{C}\boldsymbol{g}$ and $\boldsymbol{a} = \boldsymbol{F}^{+1/2}\boldsymbol{g}$. The next result measures the fraction of the unrestricted decrease that the chart retains.

Proposition 1 (Restricted decrement)**.** Over $\boldsymbol{\delta} = \boldsymbol{C}^\top \boldsymbol{z}$ the largest decrease of $\mathcal{Q}$ is $\frac{1}{2}\boldsymbol{g}_z^\top \boldsymbol{F}_z^+ \boldsymbol{g}_z = \frac{1}{2}\|\boldsymbol{\Pi}_{V_C}\boldsymbol{a}\|^2$, against $\frac{1}{2}\|\boldsymbol{a}\|^2$ unrestricted, so the chart keeps the fraction

$$\gamma_{\boldsymbol{g}}(\boldsymbol{C}) \equiv \gamma(\boldsymbol{C}) := \frac{\boldsymbol{g}_z^\top \boldsymbol{F}_z^+ \boldsymbol{g}_z}{\boldsymbol{g}^\top \boldsymbol{F}^+ \boldsymbol{g}} = \frac{\|\boldsymbol{\Pi}_{V_C}\boldsymbol{a}\|^2}{\|\boldsymbol{a}\|^2} \in [0, 1]. \tag{3}$$

To compare chosen and random spans, decompose $\boldsymbol{F} = \boldsymbol{U}_r \boldsymbol{\Lambda}_r \boldsymbol{U}_r^\top + \boldsymbol{E}$ with $\boldsymbol{U}_r^\top \boldsymbol{U}_r = \boldsymbol{I}_r$, $\boldsymbol{\Lambda}_r \succ 0$ and $\boldsymbol{E} \succeq 0$. Let $\boldsymbol{C} = \boldsymbol{R}_S \boldsymbol{\Psi}$ sample d rows of an orthogonal basis $\boldsymbol{\Psi}$, independently of the parameter state. The next quantity measures how unevenly the leading Fisher directions occupy those rows.

Definition 2 (Architecture–basis coherence)**.** $\mu_{\boldsymbol{\Psi}}(\boldsymbol{U}_r) := \frac{p}{r} \max_k \|\boldsymbol{e}_k^\top \boldsymbol{\Psi} \boldsymbol{U}_r\|_2^2 \in [1, p/r]$.

Theorem 1 (The aiming band)**.** Let $\boldsymbol{E} = 0$, and call a chart *complete* if $\gamma_{\boldsymbol{g}}(\boldsymbol{C}) = 1$ for all $\boldsymbol{g} \in \operatorname{Im} \boldsymbol{F}$. For $d < r$ no chart is complete, and some $\boldsymbol{g}$ has $\gamma_{\boldsymbol{g}} = 0$. A chart whose span contains $\operatorname{Im} \boldsymbol{U}_r$ is complete at $d = r$. A random subsampled chart is complete with probability at least $1 - \delta$ once $d \geq 4c\, \mu_{\boldsymbol{\Psi}}(\boldsymbol{U}_r)\, r \log(2r/\delta)$, by a matrix Chernoff embedding of $\operatorname{Im} \boldsymbol{U}_r$ (Proposition 4).

The rank and sampling bounds give

$$r \;\leq\; d \;<\; 4c\, \mu_{\boldsymbol{\Psi}}(\boldsymbol{U}_r)\, r \log(2r/\delta). \tag{4}$$

Below r, aiming can help particular gradients but cannot be complete. The upper edge is a sufficient bound for random completeness, not a predicted accuracy transition. It scales with the coherence: a random sign diagonal lowers $\mu_{\boldsymbol{\Psi}}$ to $O(1 + \log(p/\delta)/r)$ with high probability (Tropp, 2011), and so should narrow the band. With a spectral tail, r is a numerical rank and completeness holds up to the tail term of Proposition 4.

Geometry is scale-free: it tells whether a direction is represented, not whether its training signal vanishes near the base point. Functional order decides the latter.

3.3 VALUATION: THE ORDER AT WHICH AN OPERATION BECOMES ACCESSIBLE

To predict the order before training, take zero-centred attention: $\boldsymbol{W}_Q, \boldsymbol{W}_K = O(\varepsilon)$, so their product is $O(\varepsilon^2)$ and differentiating by either factor leaves one power; in the graph transformer the value, output and head add three more (Figure 2). The graph rule counts every such factor on each path from input to loss.

Definition 3 (Valuation graph)**.** A host is a directed acyclic graph whose sources are the data, the parameters outside the chart, and the generated tensors $\boldsymbol{W}_j = \boldsymbol{W}_j^{(0)} + \operatorname{vec}^{-1}(\boldsymbol{C}_j^\top \boldsymbol{z}_j)$, biases included. Every other node applies a multilinear map (linear layers, convolutions, $\boldsymbol{Q}\boldsymbol{K}^\top$, $\boldsymbol{P}\boldsymbol{V}$), a sum, an entrywise activation that is positively homogeneous of degree one or analytic with $\sigma(0) = 0$, a softmax, or a normalization $\boldsymbol{b} \mapsto \boldsymbol{\Pi}\boldsymbol{b}/(\operatorname{mean}((\boldsymbol{\Pi}\boldsymbol{b})^2) + \delta)^{1/2}$ with $\boldsymbol{\Pi}$ a centring projector or $\boldsymbol{I}$.

Let $v(b), \bar{v}(b)$ be graph-assigned powers of a node's value and loss gradient. A zero-base generated tensor has $\boldsymbol{W}_j^{(0)} = 0$ and starts at power one; other sources start at zero. Products add powers; sums use the minimum, a bound if leading terms cancel. Backpropagation through a product counts its other factors. Softmax removes constant shifts. Scale-invariant normalization resets the forward power to zero and subtracts the centred input power in backpropagation; a dominant stabilizer instead preserves the centred input power. The centred powers $v_\Pi, \bar{v}_\Pi$ discard leading terms constant along the normalized or softmax axis, such as broadcast biases; no other cancellation is modelled.

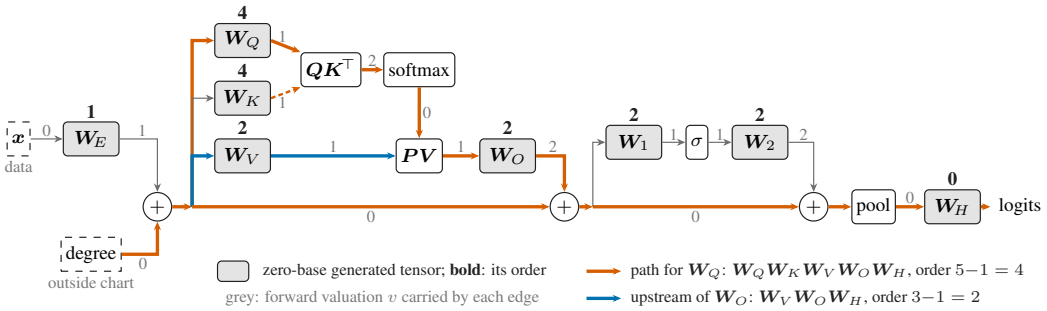


Figure 2: Power count for a graph-transformer block at a zero base point. After differentiating with respect to the query (orange), key, value, output and head supply four powers: order four. The output's path starts at the value (blue), so counting only downstream factors misses one: order two. Layer norms (forward power zero) are omitted.

With $\bar{v}(\text{logits}) = 0$ and $\bar{v}(b) = \min_{c \in \text{out}(b)} \kappa(b,c)$,

$$v(c) = \begin{cases} \sum_{b \in \text{in}(c)} v(b) & \text{multilinear} \\ \min_{b \in \text{in}(c)} v(b) & \text{sum} \\ v(b) & \text{activation} \\ 0 & \text{softmax; norm. } \delta{=}0 \\ v_\Pi(b) & \text{norm., } \delta \text{ dominant} \end{cases} \qquad \kappa(b,c) = \begin{cases} \bar{v}(c) + \sum_{b' \neq b} v(b') & \text{multilinear} \\ \bar{v}(c) - v_\Pi(b) & \text{norm. } \delta{=}0 \\ \bar{v}_\Pi(c) & \text{softmax} \\ \bar{v}(c) & \text{otherwise.} \end{cases} \tag{5}$$

These rules give the following bound.

Theorem 2 (Valuation rule)**.** Fix a finite batch and a loss with bounded, real-analytic logit gradient, and let $\mathcal{U}$ be the full-measure set of $\boldsymbol{u}$ at which the leading coefficient of $\boldsymbol{\Pi b}(\varepsilon \boldsymbol{u})$ is nonzero at every normalization with $\delta = 0$. Then for every zero-base generated $\boldsymbol{W}_\ell$ and every $\boldsymbol{u} \in \mathcal{U}$,

$$\|\nabla_{\boldsymbol{z}_\ell} \mathcal{L}(\varepsilon \boldsymbol{u})\| \leq \|\partial \mathcal{L} / \partial \boldsymbol{W}_\ell(\varepsilon \boldsymbol{u})\| = O\big(\varepsilon^{\bar{v}(\boldsymbol{W}_\ell)}\big). \tag{6}$$

If the coefficient $\hat{\boldsymbol{G}}_\ell(\boldsymbol{u})$ of $\varepsilon^{\bar{v}(\boldsymbol{W}_\ell)}$ in $\nabla_{\boldsymbol{z}_\ell} \mathcal{L}$ does not vanish identically on any open set (*non-cancellation*), then $\text{ord}_\ell(\boldsymbol{u}) = \bar{v}(\boldsymbol{W}_\ell)$ for almost every $\boldsymbol{u}$.

Under non-cancellation, the graph fixes the order before training. Between normalizations, let $q_{\ell,p}$ count zero-base generated factors multiplying a source-to-logit path p through $\boldsymbol{W}_\ell$, including $\boldsymbol{W}_\ell$ and side inputs at product nodes:

$$\text{ord}_\ell = \min_p \big(q_{\ell,p} - 1\big). \tag{7}$$

Start each path at a source, not at block ℓ (Figure 2). A latent shared across blocks receives their summed gradients, led by the lowest order. The following cases are tested below.

Corollary 1 (Instances)**.** At a base point giving every generated tensor a nonzero value, as at our hosts' initializations and in Li et al. (2018), every block has order zero under non-cancellation. At a zero base point, each score factor has one vanishing partner; freezing one factor or generating their product as one tensor removes that local power. In the graph transformer, the weights of query and key have order four, value, output and feed-forward two, the embedding one and the head zero (Figure 2); a latent that also generates a bias of lower order, as at the output and second feed-forward layers, has order one less. With dominant generated biases, L generated layers under a generated classifier have orders $1, \ldots, L$ from the classifier down and L again at the first layer. A normalization with $\delta = 0$ after every layer gives zero throughout, or -1 with the classifier outside the chart, the known batch-normalization scale invariance (van Laarhoven, 2017; Arora et al., 2019); when the stabilizer dominates, each of N generated tensors has order $N - 1$.

3.4 Geometry and valuation require different interventions

Changing the chart changes its span or size. Changing the parameterization changes which tensors are generated or their base values.

Theorem 3 (Separation)**.** (a) *Geometry sees only the span:* for invertible $\boldsymbol{D}$, charts $\boldsymbol{C}$ and $\boldsymbol{DC}$ give the same decrement and $\operatorname{rank} \boldsymbol{F}_z$. (b) *Geometry is blind to scale:* if blocks enter at orders r_b and the rescaled tangent has a limit of constant rank (Appendix A.3), the decrement has an ε-free limit along $\varepsilon \boldsymbol{u}$, while $\|\boldsymbol{g}_z\| = O(\varepsilon^m)$ and $\operatorname{tr} \boldsymbol{F}_z = O(\varepsilon^{2m})$, $m = \min_b r_b$. (c) *Valuation ignores the chart:* under the hypotheses of Theorem 2, ord_ℓ is the same for every chart satisfying non-cancellation.

At fixed base point and generated tensors, changing the chart's span can change geometry but not a block's order under non-cancellation. Changing the base point or generated tensors can change that order. Since γ is scale-free it cannot diagnose collapse, whereas gradient norms and $\operatorname{tr} \boldsymbol{F}_z$ vanish with the order.

3.5 What functional order implies for training

Functional order describes gradients near the base point; to affect training, it must delay movement. Changing only $\boldsymbol{z}(0) = \varepsilon \boldsymbol{u}$ leaves the reachable weights fixed, so its effect is dynamical rather than a change in loss-set intersection (Larsen et al., 2022; Cheng et al., 2026). Moving $\boldsymbol{\theta}_0$ or merging a product changes that set. Gradient descent sees vanishing gradients directly; normalized optimizers also depend on the product's degree. The theorem bounds both cases.

Theorem 4 (Degree and escape)**.** (a) Let a block output $\boldsymbol{Y}$ be multilinear, with bound M, in q zero-base generated tensors and in auxiliary inputs of norm at most α. If every latent step satisfies $\|\boldsymbol{z}_j(t+1) - \boldsymbol{z}_j(t)\| \le \bar{\eta}_j$, then

$$\|\boldsymbol{Y}(T)\| \le M\alpha \prod_{j=1}^{q} \big(\varepsilon \|\boldsymbol{u}_j\| + \bar{\eta}_j T\big), \tag{8}$$

and Adam with $\beta_1^2 < \beta_2$ and learning rate η has $\bar{\eta}_j \le c_A \eta \sqrt{d_j}$, with c_A depending only on (β_1, β_2). (b) For zero-base factors in the score $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top$ with $\boldsymbol{G}_S = \partial\mathcal{L}/\partial\boldsymbol{S}$, $\hat{\boldsymbol{W}}(\boldsymbol{e}_i) = \operatorname{vec}^{-1}(\boldsymbol{C}^\top \boldsymbol{e}_i)$ and a block input $\boldsymbol{X}$ that does not depend on this block's $\boldsymbol{W}_Q, \boldsymbol{W}_K$, latent gradient flow satisfies exactly

$$\frac{d}{dt}\begin{bmatrix}\boldsymbol{z}_Q \\ \boldsymbol{z}_K\end{bmatrix} = -\begin{bmatrix}0 & \boldsymbol{K}(t) \\ \boldsymbol{K}(t)^\top & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{z}_Q \\ \boldsymbol{z}_K\end{bmatrix}, \quad \boldsymbol{K}_{ij} = \mathbb{E}\,\big\langle \boldsymbol{G}_S, \boldsymbol{X}\hat{\boldsymbol{W}}_Q(\boldsymbol{e}_i)\hat{\boldsymbol{W}}_K(\boldsymbol{e}_j)^\top\boldsymbol{X}^\top\big\rangle. \tag{9}$$

So $\mathcal{M}_0 = \{\boldsymbol{z}_Q = \boldsymbol{z}_K = 0\}$ is invariant, and reaching norm ρ takes $T_{\mathrm{esc}} \ge \log(\rho/(\varepsilon\|\boldsymbol{u}_{QK}\|))/\bar{\sigma}$ with $\bar{\sigma} = \sup_{t \le T_{\mathrm{esc}}} \|\boldsymbol{K}(t)\|_{\mathrm{op}}$.

Part (b) is exact because $\partial\mathcal{L}/\partial\boldsymbol{W}_Q$ is linear in $\boldsymbol{W}_K$. Under gradient flow a positive order delays escape logarithmically in $1/\varepsilon$; under bounded-step optimizers a block grows as the q-th power of the step budget, a merged one as the first. A nonzero base can remove the invariance of $\mathcal{M}_0$: $\dot{\boldsymbol{z}}_Q$ gains the term $-\boldsymbol{C}_Q \operatorname{vec}(\boldsymbol{X}^\top\boldsymbol{G}_S\boldsymbol{X}\boldsymbol{W}_K^{(0)})$, possibly nonzero at the base point. Both parts are bounds, not trajectory predictions; suppression ends once $\varepsilon\|\boldsymbol{u}\|$ or $\bar{\eta}T$ is large enough to affect the block's output.

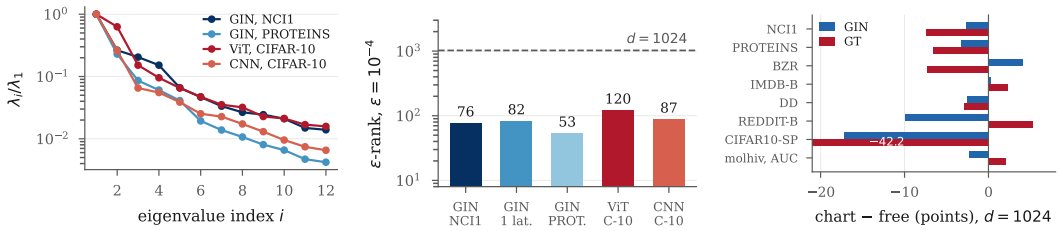
3.6 An exactly solvable case

A scalar model isolates the same mechanism without architectural detail. Fit a scalar score s to $s_\star > 0$ under $\frac{1}{2}(s_\star - s)^2$, generated in the three ways of Corollary 1: factored, $s = ab$, $a(0) = b(0) = \varepsilon$; frozen, $s = ab_0$, $a(0) = \varepsilon$; merged, $s = c$, $c(0) = 0$.

Proposition 2 (Time to half the target)**.** Let $b_0 > 0$ and $0 < \varepsilon < \min\{\sqrt{s_\star/2}, s_\star/(2b_0)\}$. Under gradient flow, and under sign flow of speed η (the continuous form of a per-coordinate normalized step), s first reaches $s_\star/2$ at

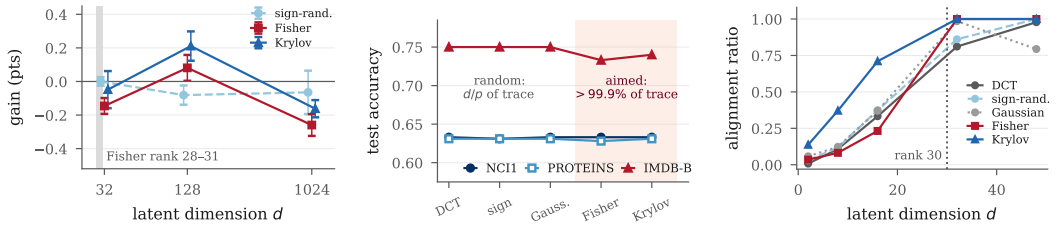
$$\Big(T_{\mathrm{fac}}, T_{\mathrm{frozen}}, T_{\mathrm{merged}}\Big) = \Big(\tfrac{\log(s_\star/\varepsilon^2 - 1)}{2s_\star},\ \tfrac{\log(2 - 2\varepsilon b_0/s_\star)}{b_0^2},\ \log 2\Big) \text{ and } \Big(\tfrac{\sqrt{s_\star/2} - \varepsilon}{\eta},\ \tfrac{s_\star/(2b_0) - \varepsilon}{\eta},\ \tfrac{s_\star}{2\eta}\Big). \tag{10}$$

Under gradient flow the factored score is logistic, $\dot{s} = 2s(s_\star - s)$, so its escape time grows as $s_\star^{-1}\log(1/\varepsilon)$ while the other two stay bounded as $\varepsilon \to 0$. Under sign flow each factor grows linearly in time and the score quadratically, so as $\varepsilon \to 0$ the factored form is slower than the merged one exactly when $s_\star < 2$ in units of the host's weight scale. It separates factorization from parameter count, consistent with singular learning theory (Watanabe, 2009).



(a) Fisher spectrum (b) Numerical rank against d (c) One budget, GIN and GT

Figure 3: **Distortion: the Fisher spectrum, numerical rank, and one budget on two hosts.**



(a) GT, NCI1 (b) GT, three benchmarks (c) $\gamma(\boldsymbol{C})$, IMDB-B

Figure 4: **Orientation at two base points, and the alignment ratio.**

4 Experiments

Test i answers Qi; Test 4 (distortion) comes first, then the collapse tests.

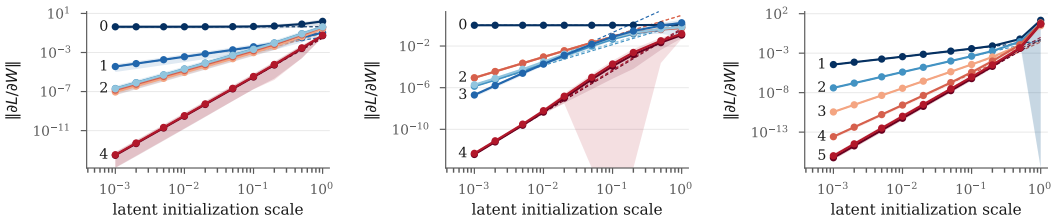
- **Q1** Do measured gradient slopes follow the integer orders read from the computation graph (Theorem 2), and does the leading term survive projection onto redrawn charts?
- **Q2** Does a shared latent (a different chart) leave those orders unchanged (Theorem 3)?
- **Q3** Does merging or freezing a factor lower an order while rescaling does not (Corollary 1), and does that ordering persist in finite-time escape (Theorem 4)?
- **Q4** Does aiming a chart help only in the intermediate regime that Theorem 1 motivates?

The base-point experiment changes both reachable weights and gradient orders, so its accuracy shift cannot be attributed to order alone. The hosts are a five-layer graph isomorphism network (Xu et al., 2019), written GIN; a graph transformer matched to it in generated-parameter count, written GT; a compact vision transformer (Dosovitskiy et al., 2021), written ViT; and a CNN matched to ViT in generated-parameter count. Each layer uses frozen rows of a discrete cosine transform (DCT) basis (Appendix D). Unless stated, $d = 1024$, the base point is zero, and accuracies average three training seeds. Figure 4a also averages five chart draws; the non-cancellation check uses ten. Appendices E, F and G give every setting and experiment in full.

4.1 Distortion: chart design matters at intermediate budgets

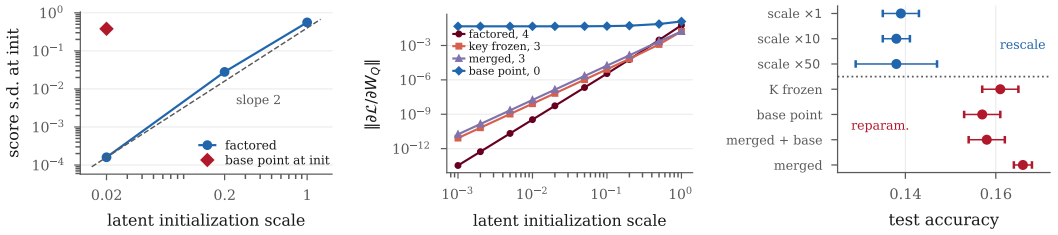
The pullback Fisher is strongly anisotropic on GIN and both image hosts (Section 3.1). At trained solutions, which are themselves chart-dependent, the pullback Fisher falls by about two decades over its first twelve eigenvalues, its ten leading directions of 1024 carry 91 to 97 percent of its trace, and its numerical rank at 10^{-4} of the largest eigenvalue is 53 to 120 (Figure 3a, b; Table 5; CIFAR-10 for the image hosts). The same budget then costs hosts different amounts: GT loses 7.4 points to the chart on NCI1 where GIN loses 2.7, and on REDDIT-B the sign flips (c; Appendix J).

Test 4: aiming helps only in an intermediate regime (Theorem 1). With the base point at the host's initialization, five independent chart draws on NCI1 give a small advantage only at $d = 128$: gradient Krylov gains $+0.21 \pm 0.09$ points over DCT rows and Fisher eigenvectors $+0.08 \pm 0.08$ (both positive in $4/5$ draws). Neither helps at $d = 32$; at $d = 1024$ both lose in all five draws (Figure 4a, Table 7). Chart-draw variability is no larger than training-seed variability in every cell. At the zero base point, the alignment ratio separates chart families before training mainly below the



(a) GT, NCI1 (b) ViT, CIFAR-100 (c) CNN, no normalization

Figure 5: **Test 1: weight-gradient norm versus latent scale, by operation.**



(a) ViT, score s.d. at init (b) GT, four parameterizations (c) ViT, CIFAR-100 accuracy

Figure 6: **Test 3: rescaling against reparameterizing the attention score.**

Fisher rank (Figure 4c, Table 8), and at $d = 1024$ no aimed or coherence-reducing family reliably beats DCT rows (b; Table 24).

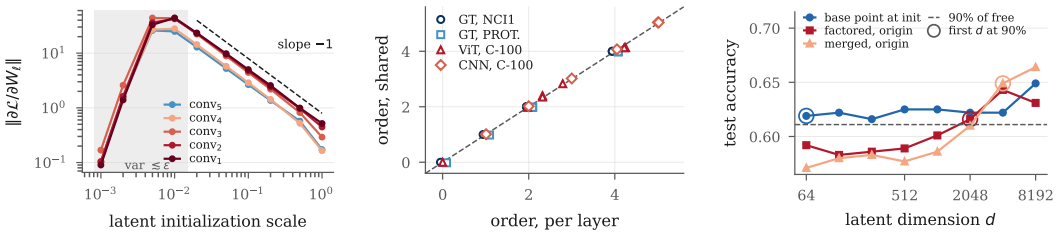
4.2 Collapse: the parameterization sets the order, and only the parameterization moves it

Test 1: gradient scaling follows the integer orders of the valuation graph (Theorem 2). We fit the slope of $\log\|\partial\mathcal{L}/\partial\boldsymbol{W}_\ell\|$ against $\log\varepsilon$. With no lower-order bias sharing its latent, this lower-bounds the order of Definition 1; it equals that order if chart projection preserves the leading term. On GT, and on CNN with its normalization removed, every generated operation sits within 0.01 of its predicted integer, with a spread of about 0.02 over three seeds by ten batches (Figure 5, Table 9); on ViT (b) the attention path approaches 2 and 4 as the scale shrinks. A normalization resets the order: with batch normalization's variance above its stabilizer, every CNN convolution fits within 0.12 of zero with the classifier in the chart and between -0.98 and -1.11 with it outside (Figure 7a, Table 17), the scale invariance known for batch normalization (van Laarhoven, 2017; Arora et al., 2019), which the rule assigns too.

Test 3: merging or freezing a generated factor lowers its order; rescaling does not (Corollary 1). Rescaling moves the score along a line of fixed slope: on ViT the score starts about two thousand times below a free initialization, near the square of the 49.3-fold gap in its factors, and grows with slope two in the latent scale (Figure 6a, Table 20). The parameterization chooses the line: on GT the query's order is 4.00 factored, 3.00 with the key frozen, 3.00 for the merged score and 0.00 with the base point at initialization (b, Table 9). These changes also alter the score's initial scale, so slopes test the order. On ViT, the factored score at fifty times the scale starts thirty times larger than the merged score but reaches 0.138 accuracy versus 0.166 (c, Table 15). This arm changes every block and is away from the origin; its accuracy gap is consistent with the order, not attributable to it.

A chart can train around an operation it cannot move. On GT, fifty times the latent scale leaves attention within three percent of uniform after training while the merged score moves more than twice as far on NCI1 and PROTEINS, yet accuracy does not track attention movement consistently across benchmarks (Table 12): GT pools over its tokens, and its other operations carry the task whether or not attention moves.

Test 2: a shared latent leaves per-operation slopes nearly unchanged (Theorem 3). A single latent for the whole host moves every per-operation slope by at most 0.05 on GT and CNN and 0.13 on ViT (Figure 7b, Table 19).



(a) CNN, batch norm. (b) Shared against per-layer latent (c) GT, PROTEINS

Figure 7: **Normalization (Test 1), a shared latent (Test 2), and the base point.**

Host	Operation	r_W	r_z	min R/med R
GT	Q	4.00	4.00	0.89
GT	K	4.00	4.00	0.94
GT	Q (K frozen)	3.00	3.00	0.84
GT	merged $QK^\top$	3.00	3.00	0.83
ViT	Q	4.03	4.03	0.97
ViT	K	4.02	4.03	0.89

Table 1: **Projection check.**

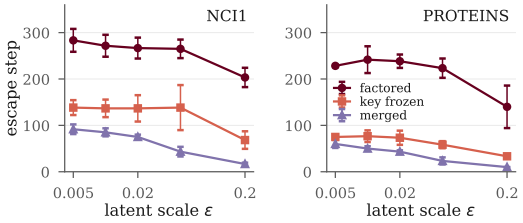


Figure 8: **Escape time of the attention score.**

The leading order survives projection and affects finite-time training. The non-cancellation condition of Theorem 2 holds empirically for the attention factors: across ten chart draws, the slopes r_W of the ambient gradient $\|\partial\mathcal{L}/\partial \boldsymbol{W}\|$ and r_z of the latent gradient $\|\nabla_{\boldsymbol{z}}\mathcal{L}\|$ against $\log\varepsilon$ agree to within 10^{-3} on GT and 0.09 on ViT, while the projected leading coefficient remains bounded away from zero: the ratio R of the two norms stays near its median across draws (Table 1). The order also has a finite-time consequence. We take the escape time as the first optimizer step at which the RMS change of the pre-softmax scores exceeds 0.01 (mean and s.d. over three seeds at five initial latent scales). Across NCI1 and PROTEINS, factored attention escapes last, freezing one factor is intermediate, and the merged score is fastest at every tested scale, the last two sharing an order and differing in initial scale (Figure 8, Table 14); merging reaches the threshold 3 to 14 times earlier than the factored form. Under Adam the factored escape time changes little as ε shrinks from 0.05 to 0.005, consistent with the bounded-step case of Theorem 4 and the sign flow of Proposition 2, rather than the logarithmic growth of gradient flow.

The base point moves the threshold crossing. With the degree embedding frozen, GT on PROTEINS first reaches the 0.611 target at $d = 64$ when centred at initialization, as in Li et al. (2018), and at 2048 when centred at zero, as in weight generators (Figure 7c, Table 16). The target is ninety percent of free accuracy with the embedding trained and 1.5 points above the 0.596 majority rate. This is one chart draw, not a precise ratio; no free baseline with the embedding frozen exists. The crossing is qualitative, not a calibrated intrinsic dimension, and we do not attribute it to order alone.

5 Conclusion

The geometry of a low-dimensional chart in weight space does not characterize the optimization problem it creates. The host decides what the optimizer sees, in two ways that call for different remedies. Distortion concentrates the chart's directions onto a few functional ones; a larger or better-oriented chart addresses it locally, and orientation mattered only at intermediate budgets. Collapse removes whole operations at first order; no change of chart removes it, but its order can be read from the computation graph before training, and changing what is generated lowers it at no cost in trainable parameters, without guaranteeing accuracy. For intrinsic-dimension-style measurements or weight generators centred at zero, this means the number describes a host–task pair at a base point rather than a task, and a low value can mean that the architecture trained around what the chart could not move. The original construction, centred at host initialization, avoids multiplicative collapse. For zero-centred weight generators, count orders before training.

AI USE STATEMENT

In this work, we used generative AI tools (a coding agent, Claude Code) to implement the training and analysis code, to give feedback on experimental design, and to help interpret results. We also used them to draft parts of the paper and polish its text, to search for related work, and to write the plotting scripts that draw the figures from the archived experimental outputs. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Anonymized code for every experiment, with the commands that run it, is available at https://anonymous.4open.science/r/Geometry-Weight-Space-ICLR2027-517B/. Appendix D specifies the chart, Appendix E gives every training setting, Appendix F lists the sampling behind each result, and Appendix A contains all proofs.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations (ICLR)*, 2023.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning (ICML)*, 2018.

Sanjeev Arora, Zhiyuan Li, and Kaifeng Lyu. Theoretical analysis of auto rate-tuning by batch normalization. In *International Conference on Learning Representations (ICLR)*, 2019.

Andrew Cheng, Ali Eslamian, Jie Cheng, Mehdi Zargham, and Qiang Cheng. Predicting when random low-dimensional reparameterizations train neural networks. *arXiv preprint arXiv:2608.12597*, 2026.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A. Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning (ICML)*, 2018.

Vijay Prakash Dwivedi, Chaitanya K. Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. *Journal of Machine Learning Research*, 24(43):1–48, 2023.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.

Stanislav Fort and Adam Scherlis. The Goldilocks zone: Towards better understanding of neural network loss landscapes. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In *International Conference on Machine Learning (ICML)*, 2019.

Frithjof Gressmann, Zach Eaton-Rosen, and Carlo Luschi. Improving neural network training in low dimensional random bases. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022.

David Ha, Andrew Dai, and Quoc V. Le. HyperNetworks. In *International Conference on Learning Representations (ICLR)*, 2017.

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open Graph Benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of Fisher information in deep neural networks: Mean field approach. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Prakhar Kaushik, Shravan Chaudhari, Ankit Vaidya, Rama Chellappa, and Alan Yuille. The universal weight subspace hypothesis. *arXiv preprint arXiv:2512.05117*, 2025.

Taeyeong Kim, Ahhyun Kim, TaeHyeon Kim, and Unggi Lee. Not the dimension, the norm: What matters in gradient-free weight perturbation of language models. *arXiv preprint arXiv:2608.01624*, 2026.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Agustinus Kristiadi, Felix Dangel, and Philipp Hennig. The geometry of neural nets' parameter spaces under reparametrization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical Fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Brett W. Larsen, Stanislav Fort, Nic Becker, and Surya Ganguli. How many degrees of freedom do we need to train deep networks: a loss landscape perspective. In *International Conference on Learning Representations (ICLR)*, 2022.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations (ICLR)*, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *International Conference on Machine Learning (ICML)*, 2015.

Ionut-Vlad Modoranu, Mher Safaryan, and Dan Alistarh. MatryoshkaLoRA: Learning accurate hierarchical low-rank representations for LLM fine-tuning. *arXiv preprint arXiv:2605.07850*, 2026.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. TUDataset: A collection of benchmark datasets for learning with graphs. In *ICML Workshop on Graph Representation Learning and Beyond*, 2020.

Weronika Ormaniec, Felix Dangel, and Sidak Pal Singh. What does it mean to be a transformer? Insights from a theoretical Hessian analysis. In *International Conference on Learning Representations (ICLR)*, 2025.

Vardan Papyan. Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet Hessians. In *International Conference on Machine Learning (ICML)*, 2019.

Paulius Rauba and Mihaela van der Schaar. Deep hierarchical learning with nested subspace networks for large language models. In *International Conference on Learning Representations (ICLR)*, 2026.

Lord Sen and Shyamapada Mukherjee. Mapping networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36215–36223, 2026.

Minhak Song, Kwangjun Ahn, and Chulhee Yun. Does SGD really happen in tiny subspaces? In *International Conference on Learning Representations (ICLR)*, 2025.

Ke Sun. Deterministic bounds and random estimates of metric tensors on neuromanifolds. In *International Conference on Learning Representations (ICLR)*, 2026.

Ke Sun and Frank Nielsen. A geometric modeling of Occam's razor in deep learning. *Information Geometry*, 8(Suppl 1):233–273, 2025. doi: 10.1007/s41884-025-00167-2.

Joel A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis*, 3(1–2):115–126, 2011. doi: 10.1142/S1793536911000787.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 3274–3287, 2023. doi: 10.18653/v1/2023.eacl-main.239.

Twan van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

Rahul Vashisht and Harish G. Ramaswamy. Faster query-key learning sharpens attention in self-attention models. In *International Conference on Machine Learning (ICML)*, 2026.

Hong Wang and Kelly Wang. Complete characterization of gauge symmetries in transformer architectures. In *NeurIPS Workshop on Symmetry and Geometry in Neural Representations (NeurReps)*, 2025.

Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*, volume 25 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, 2009. doi: 10.1017/CBO9780511800474.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.

Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Josh Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning (ICML)*, 2023.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *International Conference on Machine Learning (ICML)*, 2024.

Appendices A to C carry the proofs, the propositions the main text states without them, and a drawing of the mechanism. Appendices D and E give the chart and every training setting as the drivers run them. Appendix F says what each result supports before any of them are reported, and Appendices G to N report them. Appendix O continues the related work.

A Proofs and Technical Conditions

This appendix collects the standard reparameterization geometry that Section 3.1 uses without deriving, the proofs of every result of Section 3, the Hessian of the invariant zero-factor manifold, the preconditions attached to the alignment ratio $\gamma(\boldsymbol{C})$, the finite-population argument for random charts, and the effect of an unnormalized residual stream on the orders.

A.1 Reparameterization Geometry and the Layer-wise Fisher

Three standard facts are used in Section 3.1 without derivation.

Induced parameter dynamics. Equipping Θ with its Euclidean inner product, the generator induces the pullback metric $\boldsymbol{G}_{\mathrm{E}}(\boldsymbol{z}) = \boldsymbol{D}\boldsymbol{g}(\boldsymbol{z})^{\top}\boldsymbol{D}\boldsymbol{g}(\boldsymbol{z})$ on the latent space and the dual operator $\boldsymbol{M}(\boldsymbol{z}) := \boldsymbol{D}\boldsymbol{g}(\boldsymbol{z})\boldsymbol{D}\boldsymbol{g}(\boldsymbol{z})^{\top}$ on the ambient space. Since $\nabla_{\boldsymbol{z}}\mathcal{L} = \boldsymbol{D}\boldsymbol{g}(\boldsymbol{z})^{\top}\nabla_{\boldsymbol{\theta}}\mathcal{L}$, Euclidean gradient flow $\dot{\boldsymbol{z}} = -\nabla_{\boldsymbol{z}}\mathcal{L}$ induces $\dot{\boldsymbol{\theta}} = -\boldsymbol{M}(\boldsymbol{z})\nabla_{\boldsymbol{\theta}}\mathcal{L}$, so a mapping network acts locally as a state-dependent preconditioner and restriction of the ambient gradient. Two generators can share a latent dimension, a rank and even $\boldsymbol{M}(\boldsymbol{z})$ while producing very different functional changes in different hosts, which is why the composed tangent

$$\boldsymbol{J}_{\boldsymbol{z}}h(\boldsymbol{z};x) = \boldsymbol{J}_{\boldsymbol{\theta}}f_{\boldsymbol{\theta}}(x)\,\boldsymbol{D}\boldsymbol{g}(\boldsymbol{z}) \tag{11}$$

rather than $\boldsymbol{M}$ is the object the main text works with. The same remark applies to the neural tangent kernel of the reparameterized model, $\boldsymbol{J}_{\boldsymbol{\theta}}f_{\boldsymbol{\theta}}(x)\,\boldsymbol{M}(\boldsymbol{z})\,\boldsymbol{J}_{\boldsymbol{\theta}}f_{\boldsymbol{\theta}}(x')^{\top}$: identical parameter-space geometry induces different kernels in different hosts.

Functional rank. Stacking $\boldsymbol{J}_{\boldsymbol{\theta}}f_{\boldsymbol{\theta}}(x_i)$ over a collection $\mathcal{D} = \{x_i\}_{i=1}^{N}$ into $\mathcal{J}_{\boldsymbol{\theta}}$ and $\boldsymbol{J}_{\boldsymbol{z}}h(\boldsymbol{z};x_i)$ into $\mathcal{J}_{\boldsymbol{z}}$, Eq. (11) gives $\mathcal{J}_{\boldsymbol{z}} = \mathcal{J}_{\boldsymbol{\theta}}\boldsymbol{D}\boldsymbol{g}(\boldsymbol{z})$ and hence $\operatorname{rank}\mathcal{J}_{\boldsymbol{z}} \le \min\{\operatorname{rank}\boldsymbol{D}\boldsymbol{g}(\boldsymbol{z}), \operatorname{rank}\mathcal{J}_{\boldsymbol{\theta}}\}$: parameter-space degrees of freedom need not become independent function-space degrees of freedom. Correspondingly $\operatorname{rank}\boldsymbol{F}_{\boldsymbol{z}} \le \operatorname{rank}\boldsymbol{D}\boldsymbol{g}(\boldsymbol{z})$, strictly when $\boldsymbol{H}_f$ is singular, and on a finite $\mathcal{D}$ the finer chain $\operatorname{rank}\boldsymbol{F}_{\boldsymbol{z}} \le \operatorname{rank}\mathcal{J}_{\boldsymbol{z}} \le \operatorname{rank}\boldsymbol{D}\boldsymbol{g}(\boldsymbol{z})$ holds.

Layer-wise Fisher. Writing $h_\ell = \phi_\ell(h_{\ell-1};\boldsymbol{W}_\ell)$, $f = F_{\ell+1:L}(h_\ell)$ and $\boldsymbol{W}_\ell = g_\ell(\boldsymbol{z}_\ell)$, the tangent of block ℓ factors as

$$\boldsymbol{J}_{\boldsymbol{z}_\ell}f = \underbrace{\boldsymbol{J}_{h_\ell}F_{\ell+1:L}}_{\boldsymbol{B}_\ell}\,\underbrace{\boldsymbol{J}_{\boldsymbol{W}_\ell}\phi_\ell}_{\boldsymbol{A}_\ell}\,\boldsymbol{D}\boldsymbol{g}_\ell(\boldsymbol{z}_\ell). \tag{12}$$

Substituting Eq. (12) into Eq. (2) gives the exact identity

$$\begin{aligned}\boldsymbol{F}_{\boldsymbol{z}_\ell} &= \boldsymbol{D}\boldsymbol{g}_\ell^{\top}\,\mathbb{E}_x\big[\boldsymbol{A}_\ell(x)^{\top}\boldsymbol{B}_\ell(x)^{\top}\boldsymbol{H}_f(x)\boldsymbol{B}_\ell(x)\boldsymbol{A}_\ell(x)\big]\,\boldsymbol{D}\boldsymbol{g}_\ell,\\ \boldsymbol{v}^{\top}\boldsymbol{F}_{\boldsymbol{z}_\ell}\boldsymbol{v} &= \mathbb{E}_x\left[\left\|\boldsymbol{H}_f(x)^{1/2}\boldsymbol{B}_\ell(x)\boldsymbol{A}_\ell(x)\boldsymbol{D}\boldsymbol{g}_\ell\boldsymbol{v}\right\|_2^2\right].\end{aligned} \tag{13}$$

Only the generator differential leaves the expectation. The quadratic form on the right separates the possible causes of a small Fisher value: a latent direction may produce only a small parameter perturbation through $\boldsymbol{D}\boldsymbol{g}_\ell$, or be attenuated by $\boldsymbol{A}_\ell$, or be suppressed by $\boldsymbol{B}_\ell$, or end in a function-space direction to which the predictive distribution is insensitive. The magnitude of $\boldsymbol{F}_{\boldsymbol{z}}$ alone does not say which. At a zero base point, $\boldsymbol{A}_\ell$ and $\boldsymbol{B}_\ell$ can both vanish with ε; Theorem 2 accounts for the two together, which is why it does not need $\boldsymbol{B}_\ell$ to be non-degenerate.

A.2 Proof of Proposition 1

Write $\boldsymbol{\delta} = \boldsymbol{C}^\top \boldsymbol{z}$ and $\boldsymbol{y} = \boldsymbol{F}^{1/2}\boldsymbol{\delta}$, so that $\boldsymbol{y}$ ranges over V_C. Since $\boldsymbol{g} \in \operatorname{Im} \boldsymbol{F}$ we have $\boldsymbol{g} = \boldsymbol{F}^{1/2}\boldsymbol{a}$ with $\boldsymbol{a} = \boldsymbol{F}^{+1/2}\boldsymbol{g} \in \operatorname{Im} \boldsymbol{F}^{1/2}$, and hence $\boldsymbol{g}^\top\boldsymbol{\delta} = \boldsymbol{a}^\top\boldsymbol{y}$. The restricted model is $\mathcal{L}(\boldsymbol{\theta}_0) + \boldsymbol{a}^\top\boldsymbol{y} + \frac{1}{2}\|\boldsymbol{y}\|^2$ over $\boldsymbol{y} \in V_C$, minimized at $\boldsymbol{y} = -\boldsymbol{\Pi}_{V_C}\boldsymbol{a}$ with decrease $\frac{1}{2}\|\boldsymbol{\Pi}_{V_C}\boldsymbol{a}\|^2$. Without restriction $\boldsymbol{y}$ ranges over $\operatorname{Im} \boldsymbol{F}^{1/2} \ni \boldsymbol{a}$ and the decrease is $\frac{1}{2}\|\boldsymbol{a}\|^2 = \frac{1}{2}\boldsymbol{g}^\top\boldsymbol{F}^{+}\boldsymbol{g}$. In latent coordinates the same decrease is $\frac{1}{2}\sup_{\boldsymbol{w}}(2\boldsymbol{w}^\top\boldsymbol{g}_z - \boldsymbol{w}^\top\boldsymbol{F}_z\boldsymbol{w})$, which equals $\frac{1}{2}\boldsymbol{g}_z^\top\boldsymbol{F}_z^{+}\boldsymbol{g}_z$ because $\boldsymbol{g}_z = \boldsymbol{C}\boldsymbol{F}^{1/2}\boldsymbol{a}$ lies in $\operatorname{Im}(\boldsymbol{C}\boldsymbol{F}^{1/2}) = \operatorname{Im} \boldsymbol{F}_z$.

A.3 Proof of Theorem 3

(a) Under $\boldsymbol{C}' = \boldsymbol{D}\boldsymbol{C}$ we have $\boldsymbol{g}'_z = \boldsymbol{D}\boldsymbol{g}_z$ and $\boldsymbol{F}'_z = \boldsymbol{D}\boldsymbol{F}_z\boldsymbol{D}^\top$. For $\boldsymbol{g}_z \in \operatorname{Im} \boldsymbol{F}_z$ the variational identity $\boldsymbol{g}_z^\top\boldsymbol{F}_z^{+}\boldsymbol{g}_z = \sup_{\boldsymbol{w}}(2\boldsymbol{w}^\top\boldsymbol{g}_z - \boldsymbol{w}^\top\boldsymbol{F}_z\boldsymbol{w})$ holds, and the substitution $\boldsymbol{w} = \boldsymbol{D}^\top\boldsymbol{w}'$, a bijection of $\mathbb{R}^d$, turns the right-hand side for $\boldsymbol{C}'$ into that for $\boldsymbol{C}$. Congruence by an invertible matrix preserves rank (Sylvester's law of inertia), and conjugation by an orthogonal one preserves the spectrum. Any two charts with the same span are related by some invertible $\boldsymbol{D}$, so γ, whose denominator does not involve the chart, depends on the span alone.

(b) The conditions are as follows. With

$$\boldsymbol{\Lambda}_\varepsilon = \operatorname{blockdiag}_b(\varepsilon^{r_b}\boldsymbol{I}_{d_b}),$$

the rescaled tangent $\boldsymbol{J}_{\boldsymbol{z}}h(\varepsilon\boldsymbol{u}; x)\boldsymbol{\Lambda}_\varepsilon^{-1}$ converges, uniformly in x, to the limit $\bar{\boldsymbol{J}}(\boldsymbol{u}; x)$; the output gradient and $\boldsymbol{H}_f$ are continuous at $\varepsilon = 0$; and $\operatorname{rank} \boldsymbol{F}_z(\varepsilon\boldsymbol{u})$ is constant for small ε and equals the rank of $\bar{\boldsymbol{F}} = \mathbb{E}_x[\bar{\boldsymbol{J}}^\top\boldsymbol{H}_f\bar{\boldsymbol{J}}]$. By (a) with $\boldsymbol{D} = \boldsymbol{\Lambda}_\varepsilon^{-1}$, the restricted decrement at $\varepsilon\boldsymbol{u}$ equals the one computed from the rescaled tangent $\boldsymbol{J}_{\boldsymbol{z}}h\,\boldsymbol{\Lambda}_\varepsilon^{-1}$, that is, from $\tilde{\boldsymbol{g}} = \boldsymbol{\Lambda}_\varepsilon^{-1}\boldsymbol{g}_z$ and $\tilde{\boldsymbol{F}} = \boldsymbol{\Lambda}_\varepsilon^{-1}\boldsymbol{F}_z\boldsymbol{\Lambda}_\varepsilon^{-1}$. Uniform convergence of the rescaled tangent and continuity of the output gradient and of $\boldsymbol{H}_f$ give $\tilde{\boldsymbol{g}} \to \bar{\boldsymbol{g}}$ and $\tilde{\boldsymbol{F}} \to \bar{\boldsymbol{F}}$. The Moore–Penrose inverse is continuous along sequences of constant rank, so $\tilde{\boldsymbol{F}}^{+} \to \bar{\boldsymbol{F}}^{+}$ and the decrement converges to $\bar{\boldsymbol{g}}^\top\bar{\boldsymbol{F}}^{+}\bar{\boldsymbol{g}}$. The bounds on $\|\boldsymbol{g}_z\|$ and $\operatorname{tr} \boldsymbol{F}_z$ follow from $\boldsymbol{g}_z = \boldsymbol{\Lambda}_\varepsilon\tilde{\boldsymbol{g}}$ and $\boldsymbol{F}_z = \boldsymbol{\Lambda}_\varepsilon\tilde{\boldsymbol{F}}\boldsymbol{\Lambda}_\varepsilon$.

(c) The valuations v and $\bar{v}$ of Eq. (5) are computed from the graph, from which sources are generated and from which have zero base; no entry of $\boldsymbol{C}$ enters them. Theorem 2 identifies ord_ℓ with $\bar{v}(\boldsymbol{W}_\ell)$ for every chart satisfying non-cancellation.

A.4 Proof of Theorem 2

Fix $\boldsymbol{u} \in \mathcal{U}$ and write $c(\varepsilon)$ for the value of node c at $\boldsymbol{z} = \varepsilon\boldsymbol{u}$ on the fixed batch. We prove by induction in topological order that $c(\varepsilon) = \varepsilon^{v(c)}\hat{c} + o(\varepsilon^{v(c)})$ for some $\hat{c} = \hat{c}(\boldsymbol{u})$, possibly zero. Sources satisfy this with $\hat{\boldsymbol{W}}_j = \operatorname{vec}^{-1}(\boldsymbol{C}_j^\top\boldsymbol{u}_j)$ at zero base and with $\hat{c} = c$ otherwise.

For a multilinear node, expanding each argument and using multilinearity gives the product of leading terms at order $\sum v(b)$, and every other term is of strictly higher order. For a sum, the terms of order $\min v(b)$ collect into $\hat{c}$. For a positively homogeneous activation σ with Lipschitz constant L_σ and $v = v(b)$, $\|\sigma(\varepsilon^v(\hat{b} + o(1))) - \varepsilon^v\sigma(\hat{b})\| = \varepsilon^v\|\sigma(\hat{b} + o(1)) - \sigma(\hat{b})\| \le \varepsilon^v L_\sigma\, o(1)$. For an analytic activation with $\sigma(0) = 0$ the same holds at $v > 0$ by Taylor's theorem, and at $v = 0$ by continuity. A softmax is analytic, so its output is $\operatorname{softmax}(\hat{b}) + o(1)$ when $v(b) = 0$, and the uniform distribution plus $O(\varepsilon^{v(b)})$ when $v(b) > 0$; in both cases the order is zero. For a normalization with $\delta = 0$, the hypothesis gives $\boldsymbol{\Pi}\boldsymbol{b} = \varepsilon^{v_\Pi}(\hat{\boldsymbol{b}}_\Pi + o(1))$ with $\hat{\boldsymbol{b}}_\Pi \neq 0$, and scale invariance gives the output $\hat{\boldsymbol{b}}_\Pi/\operatorname{rms}(\hat{\boldsymbol{b}}_\Pi) + o(1)$, of order zero. In the regime where the stabilizer dominates, the map is $\boldsymbol{\Pi}\boldsymbol{b}/\sqrt{\delta}$ up to a relative error of order $\operatorname{mean}((\boldsymbol{\Pi}\boldsymbol{b})^2)/\delta$, so the order is $v_\Pi(\boldsymbol{b})$.

The backward pass is the same induction in reverse topological order on the cotangents $\bar{c} = \partial\mathcal{L}/\partial c$, starting from the logits, where $\bar{c}$ is bounded. The cotangent of b is the sum over consumers c of $\boldsymbol{J}_b c^\top \bar{c}$, and each local Jacobian has the order given by the corresponding row of Eq. (5). The Jacobian of a multilinear node with respect to one argument is the multilinear map with the remaining arguments inserted, of order $\sum_{b' \neq b} v(b')$. Sums and activations have Jacobians of order zero, since the derivative of a positively homogeneous activation is bounded. The softmax Jacobian is the centring projector scaled by $1/n$ at uniform attention, and $\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top$ otherwise; it annihilates the component of the cotangent that is constant along the softmax axis, which is the rule $\bar{v}_\Pi$. For a normalization with $\delta = 0$, differentiating $\boldsymbol{\Pi}\boldsymbol{b}/\operatorname{rms}(\boldsymbol{\Pi}\boldsymbol{b})$ gives $\operatorname{rms}(\boldsymbol{\Pi}\boldsymbol{b})^{-1}(\boldsymbol{I} - \boldsymbol{n}\boldsymbol{n}^\top)\boldsymbol{\Pi}$ with $\boldsymbol{n}$ the unit output direction, which is of order $-v_\Pi(\boldsymbol{b})$ by the hypothesis. The sum over consumers is of order at least the minimum, which gives $\|\partial\mathcal{L}/\partial\boldsymbol{W}_\ell\| = O(\varepsilon^{\bar{v}(\boldsymbol{W}_\ell)})$. The latent gradient is $\boldsymbol{C}_\ell \operatorname{vec}(\partial\mathcal{L}/\partial\boldsymbol{W}_\ell)$, and $\|\boldsymbol{C}_\ell\|_{\mathrm{op}} = 1$ gives Eq. (6).

For the second claim we build, alongside the forward induction, a family $\mathcal{C}$ of disjoint open subsets of $\mathcal{U}$ (*cells*) whose union has full measure and on each of which every leading coefficient computed so far is real analytic in $\boldsymbol{u}$. Initially $\mathcal{C} = \{\operatorname{int}\mathcal{U}\}$, and the leading coefficients of the sources are linear in $\boldsymbol{u}$. Multilinear maps, sums, analytic activations and softmax preserve analyticity on a cell, and so does a normalization with $\delta = 0$, whose denominator is nonzero on $\mathcal{U}$. At a positively homogeneous activation, replace each connected component A of each cell by the open set on which every coordinate of the leading pre-activation coefficient that is not identically zero on A has a fixed nonzero sign. The removed set lies in the zero set of finitely many analytic functions, none identically zero on A, so it is null. A coordinate that vanishes identically on A has a true valuation above the assigned one, which leaves the upper bound intact and contributes a zero leading term. On the refined cells each such activation acts on the leading term as a fixed diagonal linear map, so analyticity is preserved. An open set has countably many connected components, and the graph has finitely many nodes, so after the last node $\mathcal{C}$ is a countable family of cells covering $\mathcal{U}$ up to a null set, on each of which $\hat{\boldsymbol{G}}_\ell$ is real analytic.

A real-analytic function that does not vanish identically on a connected open set has a zero set of measure zero. Under non-cancellation, $\hat{\boldsymbol{G}}_\ell$ is not identically zero on any component of any cell, so its zero set is a countable union of null sets and hence null. For almost every $\boldsymbol{u}$, therefore, $\hat{\boldsymbol{G}}_\ell(\boldsymbol{u}) \neq 0$ and $\|\nabla_{\boldsymbol{z}_\ell}\mathcal{L}(\varepsilon\boldsymbol{u})\| = \varepsilon^{\bar{v}(\boldsymbol{W}_\ell)}\|\hat{\boldsymbol{G}}_\ell(\boldsymbol{u})\|(1 + o(1))$, so $\operatorname{ord}_\ell(\boldsymbol{u}) = \bar{v}(\boldsymbol{W}_\ell)$.

The path form. Unroll the backward recursion from $\boldsymbol{W}_\ell$ to the logits. Each multilinear node on the route adds the forward valuations of its other inputs, and each of these is itself a minimum over source paths by the forward recursion; the node reading $\boldsymbol{W}_\ell$ adds the valuation of the activation it multiplies. Collecting terms, $\bar{v}(\boldsymbol{W}_\ell)$ is the minimum over source-to-logit paths through the node reading $\boldsymbol{W}_\ell$ of the number of zero-base generated tensors multiplied into the signal, excluding $\boldsymbol{W}_\ell$ itself. Between normalizations that is Eq. (7).

The instances of Corollary 1. For a chain of generated layers $\boldsymbol{a}_k = \boldsymbol{W}_k\boldsymbol{h}_{k-1} + \boldsymbol{b}_k$, $\boldsymbol{h}_k = \sigma(\boldsymbol{a}_k)$ with generated biases and a generated classifier, the forward rule gives $v(\boldsymbol{a}_k) = \min(1 + v(\boldsymbol{h}_{k-1}), 1) = 1$ for every k, since $v(\boldsymbol{h}_0) = 0$. The backward rule gives $\bar{v}(\boldsymbol{a}_k) = \bar{v}(\boldsymbol{a}_{k+1}) + v(\boldsymbol{W}_{k+1})$, so $\bar{v}(\boldsymbol{a}_k)$ counts the generated layers above layer k, and $\bar{v}(\boldsymbol{W}_k) = \bar{v}(\boldsymbol{a}_k) + v(\boldsymbol{h}_{k-1})$ adds one for every layer except the first. With a normalization with $\delta = 0$ after every layer and the classifier outside the chart, $\bar{v}(\boldsymbol{a}_L) = 0 - v_\Pi(\boldsymbol{a}_L) = -1$ and $v(\boldsymbol{h}_k) = 0$, so $\bar{v}(\boldsymbol{W}_k) = -1$ at every layer, and one generated classifier adds one throughout. In the regime where the stabilizer dominates, the centring removes the broadcast bias from each $\boldsymbol{a}_k$, so $v_\Pi(\boldsymbol{a}_k) = k$ counts the generated layers below, the chain is homogeneous, and every generated tensor collects all the others.

A.5 What a fitted slope measures

Section 4 measures an order as a slope over a finite range of scales. Under the hypotheses of Theorem 2:

Proposition 3 (What a fitted slope measures)**.** If the next exponent in the expansion of $\nabla_{\boldsymbol{z}_\ell}\mathcal{L}$ exceeds $\bar{v}(\boldsymbol{W}_\ell)$ by $\Delta > 0$, the least-squares slope of $\log\|\nabla_{\boldsymbol{z}_\ell}\mathcal{L}\|$ against $\log\varepsilon$ on an equally spaced grid in $\log\varepsilon$ over $[\varepsilon_1, \varepsilon_2]$, with at least two points, is $\operatorname{ord}_\ell + O\big(\varepsilon_2^{\Delta}/\log(\varepsilon_2/\varepsilon_1)\big)$.

Proof. Under non-cancellation, $\log\|\nabla_{\boldsymbol{z}_\ell}\mathcal{L}(\varepsilon\boldsymbol{u})\| = \mathrm{ord}_\ell \log\varepsilon + \log\|\hat{\boldsymbol{G}}_\ell\| + R(\varepsilon)$ with $|R(\varepsilon)| = O(\varepsilon^{\Delta})$. The least-squares slope over $[\varepsilon_1, \varepsilon_2]$ in $\log\varepsilon$ equals ord_ℓ plus the slope of R, which is at most $\sup|R|$ over the interval divided by a constant multiple of $\log(\varepsilon_2/\varepsilon_1)$, that is, by $O(\varepsilon_2^{\Delta}/\log(\varepsilon_2/\varepsilon_1))$. The Fisher block is $\mathbb{E}[\boldsymbol{J}_{\boldsymbol{z}_\ell} h^\top \boldsymbol{H}_f \boldsymbol{J}_{\boldsymbol{z}_\ell} h]$; the gradient and the tangent have the same order whenever the output gradient is not orthogonal to the leading tangent, and then the block scales as $\varepsilon^{2\,\mathrm{ord}_\ell}$. □

A path whose competitor has one more factor but a large coefficient has a small effective gap and converges slowly, which is how the ViT attention orders behave.

A.6 Proof of Theorem 4

(a) With orthonormal chart rows, $\|\boldsymbol{W}_j\|_F = \|\boldsymbol{z}_j\|$. By the triangle inequality $\|\boldsymbol{z}_j(T)\| \le \varepsilon\|\boldsymbol{u}_j\| + \bar{\eta}_j T$, and multilinearity gives $\|\boldsymbol{Y}\| \le M\alpha\prod_j \|\boldsymbol{W}_j\|$. For Adam, each coordinate of the bias-corrected update is at most $c_A\eta$ in magnitude (Kingma & Ba, 2015), so the Euclidean step of a d_j-dimensional latent is at most $c_A\eta\sqrt{d_j}$.

(b) With $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top$ and $\boldsymbol{G}_S = \partial\mathcal{L}/\partial\boldsymbol{S}$, the chain rule gives, exactly,

$$\partial\mathcal{L}/\partial\boldsymbol{W}_Q = \boldsymbol{X}^\top\boldsymbol{G}_S\boldsymbol{X}\boldsymbol{W}_K, \qquad \partial\mathcal{L}/\partial\boldsymbol{W}_K = \boldsymbol{X}^\top\boldsymbol{G}_S^\top\boldsymbol{X}\boldsymbol{W}_Q,$$

because $\boldsymbol{X}$ does not depend on this block's $\boldsymbol{W}_Q, \boldsymbol{W}_K$; $\boldsymbol{G}_S$ may depend on anything, including $(\boldsymbol{z}_Q, \boldsymbol{z}_K)$. If $\boldsymbol{X}$ depended on them, as under parameter sharing, further terms would appear and the flow would no longer vanish identically on $\mathcal{M}_0$ unless they did. Substituting $\boldsymbol{W}_K = \sum_j z_{K,j}\hat{\boldsymbol{W}}_K(\boldsymbol{e}_j)$ and projecting with $\boldsymbol{C}_Q$ gives

$$\dot{z}_{Q,i} = -\textstyle\sum_j \boldsymbol{K}_{ij} z_{K,j}, \qquad \boldsymbol{K}_{ij} = \mathrm{tr}\left(\hat{\boldsymbol{W}}_Q(\boldsymbol{e}_i)^\top\boldsymbol{X}^\top\boldsymbol{G}_S\boldsymbol{X}\hat{\boldsymbol{W}}_K(\boldsymbol{e}_j)\right),$$

which is the stated inner product, and symmetrically for $\boldsymbol{z}_K$. The right-hand side vanishes on $\mathcal{M}_0$, so $\mathcal{M}_0$ is invariant. Writing $\boldsymbol{w} = (\boldsymbol{z}_Q, \boldsymbol{z}_K)$ and $\boldsymbol{H}(t)$ for the block matrix, $\frac{d}{dt}\|\boldsymbol{w}\|^2 = -2\boldsymbol{w}^\top\boldsymbol{H}\boldsymbol{w} \le 2\|\boldsymbol{H}\|_{\mathrm{op}}\|\boldsymbol{w}\|^2 = 2\|\boldsymbol{K}\|_{\mathrm{op}}\|\boldsymbol{w}\|^2$, and Grönwall's inequality gives $\|\boldsymbol{w}(t)\| \le \|\boldsymbol{w}(0)\|e^{\bar{\sigma}t}$, which rearranges to the escape-time bound of Theorem 4(b).

A.7 Proof of Theorem 1

By Proposition 1, $\gamma_{\boldsymbol{g}}(\boldsymbol{C}) = \|\boldsymbol{\Pi}_{V_C}\boldsymbol{a}\|^2/\|\boldsymbol{a}\|^2$ with $\boldsymbol{a}$ ranging over $\mathrm{Im}\,\boldsymbol{F}^{1/2} = \mathrm{Im}\,\boldsymbol{U}_r$ as $\boldsymbol{g}$ ranges over $\mathrm{Im}\,\boldsymbol{F}$. If $d < r$, then $\dim V_C \le d < r$, so $\mathrm{Im}\,\boldsymbol{U}_r \cap V_C^{\perp}$ contains a nonzero $\boldsymbol{a}$, and the corresponding $\boldsymbol{g}$ has $\gamma_{\boldsymbol{g}} = 0$. If $\mathrm{Im}\,\boldsymbol{C}^\top \supseteq \mathrm{Im}\,\boldsymbol{U}_r$, then $V_C \supseteq \boldsymbol{F}^{1/2}\,\mathrm{Im}\,\boldsymbol{U}_r = \mathrm{Im}\,\boldsymbol{U}_r$ and every $\gamma_{\boldsymbol{g}} = 1$, which is possible with $d = r$. For the random chart, $V_C = \boldsymbol{U}_r\boldsymbol{\Lambda}_r^{1/2}\,\mathrm{Im}(\boldsymbol{U}_r^\top\boldsymbol{C}^\top)$. At $\epsilon = \frac{1}{2}$, Proposition 4 gives $\boldsymbol{U}_r^\top\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{U}_r \succeq \frac{d}{2p}\boldsymbol{I}_r$ with probability at least $1-\delta$ once $d \ge 4c\,\mu_{\boldsymbol{\Psi}} r\log(2r/\delta)$, so $\boldsymbol{U}_r^\top\boldsymbol{C}^\top$ has rank r, $\mathrm{Im}(\boldsymbol{U}_r^\top\boldsymbol{C}^\top) = \mathbb{R}^r$, and $V_C = \mathrm{Im}\,\boldsymbol{U}_r$.

A.8 Proof of Proposition 2

Gradient flow. For $s = ab$, $\dot{a} = b(s_\star - ab)$ and $\dot{b} = a(s_\star - ab)$, so $\frac{d}{dt}(a^2 - b^2) = 0$ and the balanced initialization stays balanced. Then $s = a^2$ satisfies $\dot{s} = 2a\dot{a} = 2s(s_\star - s)$, whose solution is $s(t) = s_\star/\big(1 + (s_\star/\varepsilon^2 - 1)e^{-2s_\star t}\big)$; setting $s = s_\star/2$ gives T_{fac}. For $s = ab_0$, $\dot{s} = b_0\dot{a} = b_0^2(s_\star - s)$ with $s(0) = \varepsilon b_0$, so $s_\star - s(t) = (s_\star - \varepsilon b_0)e^{-b_0^2 t}$ and $s = s_\star/2$ exactly at $t = b_0^{-2}\log(2 - 2\varepsilon b_0/s_\star)$. For $s = c$ with $c(0) = 0$, $\dot{s} = s_\star - s$ gives $\log 2$ exactly.

Sign flow. With $\dot{a} = \eta\,\mathrm{sign}(\partial_a(-\mathcal{L}))$ and likewise for b, all signs are positive while $0 < s < s_\star$. Hence $a = b = \varepsilon + \eta t$ and $s = (\varepsilon + \eta t)^2$ in the factored case, $s = b_0(\varepsilon + \eta t)$ with one factor frozen, and $s = \eta t$ when merged. Solving $s = s_\star/2$ gives Eq. (10).

A.9 The Cross-Hessian on the Invariant Manifold

For the zero-base bilinear parameterization of Corollary 1, write $\boldsymbol{g}_S := \partial\mathcal{L}/\partial\boldsymbol{S}|_{\mathcal{M}_0}$ for the upstream loss derivative with respect to the score matrix. Since $\boldsymbol{S}$ is exactly bilinear in $(\boldsymbol{z}_Q, \boldsymbol{z}_K)$, its second

derivative in those coordinates has no diagonal part, and the mixed partial is the constant bilinear form

$$\delta \boldsymbol{z}_Q^\top \boldsymbol{K}\, \delta \boldsymbol{z}_K := \left\langle \boldsymbol{g}_S,\ \boldsymbol{X}(\boldsymbol{C}_Q^\top \delta \boldsymbol{z}_Q)(\boldsymbol{C}_K^\top \delta \boldsymbol{z}_K)^\top \boldsymbol{X}^\top \right\rangle, \tag{14}$$

so that the Hessian in $(\boldsymbol{z}_Q, \boldsymbol{z}_K)$ on $\mathcal{M}_0$ is the block matrix of Eq. (9). The Gauss–Newton part contributes nothing on $\mathcal{M}_0$ because it factors through the Jacobians of $\boldsymbol{S}$ in $\boldsymbol{z}_Q$ and in $\boldsymbol{z}_K$, each of which retains the other factor and so vanishes there; what remains is the residual-weighted term, which does not. The cross terms between $(\boldsymbol{z}_Q, \boldsymbol{z}_K)$ and any other latent block also vanish on $\mathcal{M}_0$, since every such derivative retains a factor of $\boldsymbol{W}_Q$ or $\boldsymbol{W}_K$. A matrix of the form $\left[\begin{smallmatrix} 0 & \boldsymbol{K} \\ \boldsymbol{K}^\top & 0 \end{smallmatrix}\right]$ has eigenvalues $\pm\sigma_i(\boldsymbol{K})$ together with zeros, so the spectrum is symmetric about the origin for any $\boldsymbol{K}$: the instability requires only $\boldsymbol{K} \neq 0$ and no sign condition on the loss.

A.10 Preconditions for the Alignment Ratio

Projection form of the residual. Writing $\boldsymbol{\Pi}_C^{(F)} = \mathrm{Proj}_{\boldsymbol{F}^{1/2}\,\mathrm{Im}(\boldsymbol{C}^\top)}$ for the Fisher-weighted projection onto the generated subspace and $\boldsymbol{\Delta} = -\boldsymbol{F}^+\boldsymbol{g}$ for the unrestricted Gauss–Newton step, the residual left by Proposition 1, $\mathcal{Q}_{\mathrm{res}} = \frac{1}{2}(1-\gamma(\boldsymbol{C}))\,\boldsymbol{g}^\top \boldsymbol{F}^+ \boldsymbol{g}$, can equivalently be written

$$\mathcal{Q}_{\mathrm{res}} = \tfrac{1}{2}\left\| (\boldsymbol{I} - \boldsymbol{\Pi}_C^{(F)}) \boldsymbol{F}^{1/2} \boldsymbol{\Delta} \right\|^2, \tag{15}$$

which says geometrically that the chart retains only the component of the required displacement lying in the transformed subspace. The corresponding restricted minimizer is $\boldsymbol{z}^\star = -(\boldsymbol{C}\boldsymbol{F}\boldsymbol{C}^\top)^+ \boldsymbol{C}\boldsymbol{g}$.

Preconditions. The bound $\gamma(\boldsymbol{C}) \in [0,1]$ in Proposition 1 requires $\boldsymbol{g} \in \mathrm{Im}\,\boldsymbol{F}$. Dropping it, the ratio can exceed one: take $\boldsymbol{F} = \mathrm{diag}(1,0)$, $\boldsymbol{g} = (1,1)^\top$ and $\boldsymbol{C} = (1,1)/\sqrt{2}$, for which $\boldsymbol{g}^\top \boldsymbol{F}^+ \boldsymbol{g} = 1$ while $\boldsymbol{C}\boldsymbol{g} = \sqrt{2}$ and $\boldsymbol{C}\boldsymbol{F}\boldsymbol{C}^\top = 1/2$, giving $\gamma(\boldsymbol{C}) = 4$.

For softmax cross-entropy the precondition holds when $\boldsymbol{g}$ and $\boldsymbol{F}$ are formed on the same inputs. Each per-example gradient is $\boldsymbol{J}_i^\top(\boldsymbol{p}_i - \boldsymbol{e}_{y_i})$, and the output-space Fisher is $\boldsymbol{H}_f(x_i) = \mathrm{diag}(\boldsymbol{p}_i) - \boldsymbol{p}_i\boldsymbol{p}_i^\top$, whose kernel is spanned by the all-ones vector. Since the components of $\boldsymbol{p}_i - \boldsymbol{e}_{y_i}$ sum to zero, that residual is orthogonal to the kernel and therefore lies in $\mathrm{Im}\,\boldsymbol{H}_f(x_i)$, so each per-example gradient lies in the image of the corresponding per-example Fisher and the average lies in $\mathrm{Im}\,\boldsymbol{F}$.

Two practical caveats follow. The argument requires $\boldsymbol{F}$ to be the analytic expectation $\mathbb{E}_x[\boldsymbol{J}^\top \boldsymbol{H}_f(x) \boldsymbol{J}]$; a Monte Carlo true-Fisher estimate formed by resampling labels from $p_{\boldsymbol{\theta}}$ has a different image in general, and the containment can fail. And weight decay adds a term $\lambda\boldsymbol{\theta}$ to $\boldsymbol{g}$ that need not lie in $\mathrm{Im}\,\boldsymbol{F}$, so $\gamma(\boldsymbol{C})$ should be computed from the data-fit gradient alone when the bound is to be interpreted as a fraction.

A.11 Random Charts

Let $\{\boldsymbol{\psi}_k\}_{k=1}^p$ be a fixed orthonormal basis of $\mathbb{R}^p$ and let S be a uniformly random subset of indices of size d, so that $\boldsymbol{P}_C = \sum_{k\in S} \boldsymbol{\psi}_k \boldsymbol{\psi}_k^\top$. At a parameter state $\boldsymbol{\theta}$ fixed before conditioning on the sampled subspace, each index is included with marginal probability d/p, whence

$$\mathbb{E}[\boldsymbol{P}_C] = \sum_{k=1}^{p} \Pr(k \in S)\, \boldsymbol{\psi}_k \boldsymbol{\psi}_k^\top = \frac{d}{p} \sum_{k=1}^{p} \boldsymbol{\psi}_k \boldsymbol{\psi}_k^\top = \frac{d}{p} \boldsymbol{I}_p,$$

using $\sum_k \boldsymbol{\psi}_k \boldsymbol{\psi}_k^\top = \boldsymbol{I}_p$, and linearity of the trace gives

$$\mathbb{E}[\mathrm{tr}(\boldsymbol{P}_C \boldsymbol{F}_{\boldsymbol{\theta}})] = \tfrac{d}{p}\,\mathrm{tr}(\boldsymbol{F}_{\boldsymbol{\theta}}). \tag{16}$$

The argument uses only orthonormality of the basis and symmetry of inclusion across indices; it therefore covers the DCT construction used here, whose sampling law is not rotationally invariant, as well as the Haar-uniform case. The conditioning caveat matters: once a chart has been used for training, $\boldsymbol{\theta}$ depends on $\boldsymbol{C}$ and the expectation no longer factors.

A.12 The active-landscape embedding

Theorem 1 rests on the following embedding, with the notation of Section 3.2 and a universal constant c.

Proposition 4 (Active-landscape embedding)**.** If $d \geq c\,\mu_{\boldsymbol{\Psi}}(\boldsymbol{U}_r)\, r \log(2r/\delta)/\epsilon^2$, then with probability at least $1-\delta$ the matrix $\frac{p}{d}\boldsymbol{U}_r^\top \boldsymbol{C}^\top \boldsymbol{C}\,\boldsymbol{U}_r$ has spectrum in $[1-\epsilon, 1+\epsilon]$, and $\lambda_i(\boldsymbol{C}\boldsymbol{F}\boldsymbol{C}^\top)$ lies within $[(1-\epsilon)\frac{d}{p}\lambda_i,\ (1+\epsilon)\frac{d}{p}\lambda_i + \|\boldsymbol{E}\|_{\mathrm{op}}]$ for $i \leq r$.

Proof.

Write $\boldsymbol{P}_C := \boldsymbol{C}^\top\boldsymbol{C}$, set $\boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{U}_r$ and let $\boldsymbol{x}_k := \boldsymbol{V}_{k:}^\top \in \mathbb{R}^r$. Orthogonality of $\boldsymbol{\Psi}$ and $\boldsymbol{U}_r$ gives

$$\sum_{k=1}^{p} \boldsymbol{x}_k \boldsymbol{x}_k^\top = \boldsymbol{I}_r, \qquad \|\boldsymbol{x}_k\|_2^2 \leq \frac{\mu_{\boldsymbol{\Psi}}(\boldsymbol{U}_r) r}{p}. \tag{17}$$

For a uniformly sampled subset S of size d,

$$\frac{p}{d}\boldsymbol{U}_r^\top \boldsymbol{P}_C \boldsymbol{U}_r = \frac{p}{d}\sum_{k\in S} \boldsymbol{x}_k \boldsymbol{x}_k^\top, \qquad \mathbb{E}\left[\frac{p}{d}\sum_{k\in S} \boldsymbol{x}_k \boldsymbol{x}_k^\top\right] = \boldsymbol{I}_r. \tag{18}$$

Matrix Chernoff concentration for sampling without replacement, together with the uniform bound in Eq. (17), yields for a universal $c_0 > 0$ and $0 < \epsilon < 1$,

$$\Pr\left(\left\|\frac{p}{d}\boldsymbol{U}_r^\top \boldsymbol{P}_C \boldsymbol{U}_r - \boldsymbol{I}_r\right\|_{\mathrm{op}} > \epsilon\right) \leq 2r \exp\left(-c_0 \frac{\epsilon^2 d}{\mu_{\boldsymbol{\Psi}}(\boldsymbol{U}_r) r}\right). \tag{19}$$

Sampling without replacement is no less concentrated than independent sampling with replacement for this positive-semidefinite sum. Choosing $d \geq c\mu_{\boldsymbol{\Psi}}(\boldsymbol{U}_r) r \log(2r/\delta)/\epsilon^2$ makes the right-hand side at most δ and proves the first claim of Proposition 4.

For the spectral consequence, write $\boldsymbol{F}_r = \boldsymbol{U}_r \boldsymbol{\Lambda}_r \boldsymbol{U}_r^\top$. The nonzero eigenvalues of $\boldsymbol{C}\boldsymbol{F}_r\boldsymbol{C}^\top$ equal those of

$$\boldsymbol{\Lambda}_r^{1/2} \boldsymbol{U}_r^\top \boldsymbol{P}_C \boldsymbol{U}_r \boldsymbol{\Lambda}_r^{1/2}. \tag{20}$$

Congruencing that first claim by $\boldsymbol{\Lambda}_r^{1/2}$ and applying eigenvalue monotonicity gives the two multiplicative bounds for $\boldsymbol{F}_r$. Finally, $\boldsymbol{C}\boldsymbol{E}\boldsymbol{C}^\top \succeq 0$ and row orthonormality implies $\|\boldsymbol{C}\boldsymbol{E}\boldsymbol{C}^\top\|_{\mathrm{op}} \leq \|\boldsymbol{E}\|_{\mathrm{op}}$; Weyl's inequality supplies the additive upper-tail term in the second claim.

A.13 Normalization Boundaries and the Residual Stream

Between the two extremes of Theorem 2 (a generated product consumed by a normalization with $\delta = 0$, which resets the valuation to zero, and one feeding $\mathrm{softmax}$ directly, which retains its multiplicative order) sits a generated projection that re-enters an unnormalized residual stream, such as $\boldsymbol{W}_V\boldsymbol{W}_O$. Its own tangent decays with the number of its generated factors, as the multilinear row of Eq. (5) requires. But the layer output is a sum of that branch and an $O(1)$ residual contribution carried from outside the chart, rather than the argument of a scale-sensitive nonlinearity, so the attenuation reduces the branch's relative contribution without driving the layer's output to a degenerate configuration. This is why the order of a block and the functional consequence of that order must be read separately, and why Theorem 2 assigns orders block by block rather than to an architecture as a whole.

B Additional Propositions and Caveats

Proposition 5 (Coherence ceiling)**.** For any unit v, $\|P_\Omega v\|^2 = \sum_{k\in S} |\langle \psi_k, v\rangle|^2 \leq d\,\mu(v)^2$ with $\mu(v) = \max_k |\langle \psi_k, v\rangle|$. If v is s-sparse in parameter coordinates and Ψ is a DCT basis then $\mu(v)^2 \leq 2s/P$, so $\|P_\Omega v\|^2 \leq 2sd/P$.

A band-limited chart therefore cannot align with a coordinate-localized curvature direction by more than a factor $2s$ over the random baseline. This is an uncertainty principle, not a conjecture, and it is the honest reason a frozen DCT chart has little alignment headroom. It also disposes of a tempting but wrong diagnosis. A random chart does not fail by missing the leading curvature directions: for a target subspace of dimension of order ten and $d = 1024$, a random chart is non-degenerately aligned with every one of them (Kumar et al., 2022, Lemma A.16). Whatever goes wrong is not a miss.

Proposition 6 (Interlacing)**.** Row-orthonormality of Ω gives Cauchy interlacing, $\lambda_i(F) \geq \lambda_i(G) \geq \lambda_{i+P-d}(F)$, hence $d_{\text{eff}}(G;\kappa) \leq d_{\text{eff}}(F;\kappa)$ for $d_{\text{eff}}(\kappa) = \sum_i \lambda_i/(\lambda_i + \kappa)$.

Interlacing bounds individual eigenvalues and d_{eff} at fixed κ; it does not order the condition number, participation ratio, top-ten mass or relative-threshold rank of the pulled-back and ambient Fishers, so the quantities of Section 4.1 describe the pullback only. We also correct how such measurements are often reported: the participation ratio $(\sum \lambda)^2/\sum \lambda^2$ is trace-dominated and is not d_{eff}; a spectrum with participation ratio 3 can have d_{eff} in the thousands.

The capture law's variance term. For completeness, with the notation of Eq. (16), $\mathrm{Var}(\mathrm{tr}\, G) = d\,\frac{P-d}{P-1}\,\mathrm{Var}_{\text{pop}}(\{C_{kk}\})$, the finite-population sampling variance of the Fisher's diagonal in the chart basis; a realised capture far below the mean therefore certifies a large spread, that is, strong coherence of the Fisher with the basis.

How large must d be? There are now two answers to different questions. The threshold analysis of Larsen et al. (2022) concerns global intersection with a loss sublevel set under a Haar-uniform affine slice; for a quadratic well its effective dimension reduces to $\sum_i \lambda_i/(\lambda_i + \kappa)$ with $\kappa = 2\epsilon/R^2$ and $R = \|\theta_0 - \theta^*\|$. Proposition 4 instead gives a local sufficient dimension for our subsampled-basis chart, scaling as $\mu_{\Psi}(\boldsymbol{U}_r) r \log r$ up to accuracy and confidence factors. The first concerns whether a random slice reaches a low-loss set; the second concerns whether one realized structured chart preserves a specified active local subspace. Neither can be substituted for the other.

What these do not imply. None of the local propositions bounds generalization or guarantees that stochastic gradient descent reaches $\mathcal{L}_{\min}$. The embedding result is a sufficient local geometric condition, not a necessary training threshold, and the exact sublevel-set identity does not license an escape-theorem bound for our Ω: those bounds require a subspace uniform on the Grassmannian, whereas a subsampled DCT basis is not.

C The Mechanism at a Glance

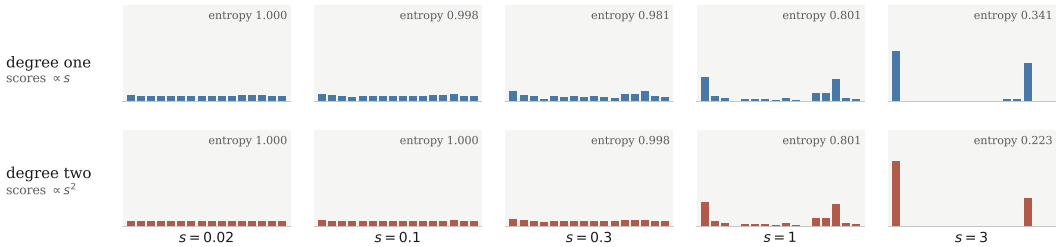


Figure 9: The square law as a filmstrip: one fixed softmax row as the generated-weight scale s grows, scores scaling as s (top) against s^2 (bottom), with each frame's normalized entropy. The profiles are synthetic; the law is the measured one of Table 20.

D Implementation of the Chart

We follow the faithful weight-generation instance of Sen & Mukherjee (2026). Each generated group of parameters is produced as $\sigma(\Omega^\top z + \alpha\|z\|^2 + b)$ with σ the identity, $\alpha = 10^{-5}$, and b initialized to zero. The scalar modulation adds a rank-one term $2\alpha z\,(\mathbf{1}^\top g)$ to the latent gradient, so the implemented chart is affine only up to it; measured at the origin and at trained latents on both graph benchmarks, its relative contribution to the latent gradient never exceeds 7×10^{-5}, so the affine analysis applies to the construction as implemented. Rows of Ω are orthonormal, drawn either as an explicit random orthogonal block when $d \times P$ fits in memory or as DCT rows with a fixed random frequency subset otherwise. Two chart variants are used throughout: a single latent for the whole network, whose Jacobian couples all layers, and one latent per generated layer, whose Jacobian is block diagonal, its budget split across layers by parameter count.

Hosts. The graph host is a five-layer graph isomorphism network with hidden width 128 (Xu et al., 2019); GT matches it in generated-parameter count with four blocks and four heads at the same

width. The image hosts are a compact ViT with four blocks, four heads, width 128, patch size four (Dosovitskiy et al., 2021), and a five-block CNN at matched generated-parameter count. The published host of Sen & Mukherjee (2026) is used unchanged for the faithfulness comparison, and its adaptation to natural images changes only the input channel count and flattened width. Normalization affines, embeddings, and GIN's epsilon parameters stay outside the chart and are counted separately in all compression ratios. Because a classifier head has few parameters, its share of a proportional budget can be a single chart row, and a single DCT row can make the head's class rows nearly identical, in which case the logit differences cannot depend on the input. The charts of the score-block and origin experiments are checked before training for the contrast between the head's class rows, and the chart seed is chosen so that the contrast is nonzero. The sweep driver does not carry that check, so the charts behind Table 23 were verified against the same criterion after the fact.

E THE TRAINING PROTOCOL AS IMPLEMENTED

This appendix reports the protocol as the drivers run it rather than as it was planned, read back from the code and from the job scripts that produced each reported number. Table 2 covers the hosts, the data and the optimization; Table 3 covers what is reported and how the chart is built. The graph benchmarks are NCI1, PROTEINS, BZR, DD, IMDB-B and REDDIT-B from TUDataset (Morris et al., 2020), the CIFAR-10 superpixel benchmark CIFAR10-SP (Dwivedi et al., 2023) and ogbg-molhiv (Hu et al., 2020); the image benchmarks are CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), MNIST (LeCun et al., 1998) and Fashion-MNIST (Xiao et al., 2017). The image hosts are trained with AdamW (Loshchilov & Hutter, 2019).

Table 2: Hosts, data and optimization, as implemented. The graph column covers both graph hosts and both graph drivers, the image column ViT and CNN, and the last column the host of Sen & Mukherjee (2026) in our replication. A dash marks a setting that family does not have.

	Graph	**Image**	**Published host**
Hosts and data			
Driver	`gt_compress`, `gt_rankctl`	`vit_cv`	`mapnet_cnn2`
Hosts	GIN, 5 layers, width 128; GT, 4 blocks, 4 heads, width 128	ViT, 4 blocks, 4 heads, width 128, patch 4; CNN, 5 convolutions	CNN2, 2 convolutions and 2 dense layers
Normalization	batch norm. in GIN, layer norm. in GT	batch norm. in the CNN, layer norm. in the ViT	none
Benchmarks	NCI1, PROTEINS, BZR, DD, IMDB-B, REDDIT-B, CIFAR10-SP, molhiv	CIFAR-10, CIFAR-100	MNIST, Fashion-MNIST, CIFAR-10, CIFAR-100
Split	fixed 90/10 for TU, official otherwise	official	
Optimization			
Optimizer	Adam	AdamW	Adam
Learning rate	10^{-3}		$3 \cdot 10^{-3}$
Schedule	cosine, per epoch	10-epoch warmup, then cosine, per step	none
Weight decay	0	0.05, never on latents or 1-D tensors	0
Batch size	128; 8 for GT on DD and REDDIT-B	256	128
Epochs	100; 30 on CIFAR10-SP and 50 on molhiv	200	20
Loss	cross entropy; binary on molhiv	cross entropy, label smoothing 0.1	cross entropy
Augmentation	—	RandAugment $(2, 9)$, crop, flip	—
Dropout	0.5 in GIN, none in GT	none	

Table 3: What is reported, and how the chart is built. The graph and image hosts of Table 2 agree on every row, so they share a column. The chart block has no counterpart in free training.

	Graph and image	**Published host**
Reporting		
Model selection	none, and no validation fold	the same
Reported number	score on the test split after the last epoch, accuracy except on molhiv	accuracy after the last epoch
Seeds	$0, 1, 2$	$1, 2, 3$
Seed controls	host and latent initialization, batch order	the same, and the chart draw
Chart		
Basis	DCT, frequencies from a seeded generator	random orthogonal block
Across seeds	identical	redrawn every seed
Latent s.d.	0.02	0.1; 0.02 on CIFAR
Allocation	by parameter count	by parameter count; equal on CIFAR
Modulation	10^{-5}	10^{-5}; smaller on CIFAR
Per-arm tuning	none	the three rows above, on CIFAR

Nothing distinguishes a chart run from a free run of the same host except the parameterization itself: the driver builds the host with its generator attached, the latent vector becomes the only trainable tensor outside the normalization affines and the embeddings, and every setting in Table 2 is inherited unchanged. No hyperparameter was selected per arm anywhere except in the published-host replication on the natural-image column, where the modulation scale, the latent initialization and the budget allocation were chosen per arm. A seed controls the host's own initialization, the latent initialization and the order in which batches are drawn. On the DCT charts it does not control the chart, which is fixed by a seed of its own and shared across runs; in the published-host replication it does. Figure 4a and Tables 7 and 10 vary the chart seed separately.

F Evidence Inventory

Two cells are reported as measured at fewer seeds than the rest and are marked in the text. The zero-base-point arm of Table 6 was run at three seeds rather than ten, because that arm carries a null result and the ten-seed extension was cut by a wall clock; and the alignment ratio of Table 8 is a single seed, because it is a quantity computed from the chart and the Fisher estimate before any training and has no seed to average over beyond the chart draw.

Four apparent effects we encountered during this study are not reported as results, because they were artifacts: GT's collapse to a single class under the frozen chart, which was a degenerate classifier chart seeded by target index; the collapse of the local dimension along an interpolant, which was batch-normalization statistics re-estimated with an exponential moving average on a host whose activation variance is far below one; the constant predictor returned by the aimed families at the larger budget with the base point moved, which was a Lanczos breakdown test that could not see a non-finite residual; and the four-point advantage of aiming on PROTEINS, which was three seeds on a benchmark whose seed spread is four points. The second and fourth are described in Appendices K and G.2.

G The Experiments in Full

Section 4 plots the experiments below in Figures 3 to 8; this appendix gives each of them in full, with the tables the panels are drawn from.

Protocol. The hosts are those of Appendix D: a five-layer graph isomorphism network (Xu et al., 2019), the message-passing host, written GIN in tables; a graph transformer matched to it in generated-parameter count, with four blocks and four heads, written GT; a compact vision transformer (Dosovitskiy et al., 2021), written ViT; and a convolutional network matched to it, written CNN. The chart is the frozen linear map of Section 3.1, one latent per generated layer with the bud-

Table 4: Each main result of this paper, with the sampling behind it and a plain verdict on whether it carries the claim it is attached to. The single entry marked RETIRED is reported here and withdrawn in Appendix G.2.

	Claim	Sampling	Verdict
1	Pulled-back metric is severely anisotropic; local dimension comparable to width	3 hosts, two estimators, at trained solutions	carries
2	Block weight-gradient slopes match the predicted integer exponents	GT and CNN without normalization; ViT's attention path approaching, 3 seeds $\times$ 10 batches, s.d. about 0.02	carries
3	The chart does not cancel the leading term of the attention factors	GT and ViT, 10 chart draws, 2 seeds $\times$ 3 batches; CNN in the text	carries
4	The order belongs to an operation's position, not its name	same run; graph embed 1 against vision patch embed 3	carries
5	The divergence regime: order -1 when the classifier is outside the chart	GIN on NCI1 and CNN on CIFAR-100, 3 seeds $\times$ 10 batches	carries
6	Attention moves in the order the analysis assigns	3 benchmarks, 3 seeds $\times$ 10 batches, paired per batch	carries
7	The factored score moves last and the merged score first, in a time that changes little as the latent scale shrinks tenfold	GT, 2 benchmarks, 5 latent scales, 3 seeds, one chart draw	carries
8	On GT the three parameterizations do not separate on accuracy	3 benchmarks, 3 seeds	carries
9	On ViT they separate by 2.7 points	CIFAR-100, 3 seeds	carries
10	Base point and latent scale are different moves	vision host, 3 seeds; graph host does not separate them	carries
11	Removing normalization collapses the chart	2 hosts, 4 benchmarks, 3 seeds	carries
12	The order reset is due to rescaling alone	both hosts, 3 seeds $\times$ 10 batches, with an ϵ sweep	carries
13	Trace capture is not alignment	5 chart families, 2 benchmarks	carries
14	Aiming does not help at the paper's budget, at either base point	both benchmarks; init base 10 seeds, zero base 3 seeds	carries
15	Aiming helps slightly at an intermediate budget once the base point is moved	NCI1, 5 draws $\times$ 3 seeds	carries
15b	The same effect on PROTEINS	10 seeds; that benchmark cannot resolve it	RETIRED
16	The alignment ratio discriminates mainly below the Fisher's numerical rank, and orders the families by task awareness rather than by trace capture	5 families, budgets from 2 to 48, one seed, computed before training	carries
17	The chart costs image hosts about fifty points	CIFAR-100, 3 seeds	carries

get split by parameter count, and its matrix, $\boldsymbol{C}$ here and Ω in the appendices, has rows that are DCT vectors at a random subset of frequencies. Unless stated, $d = 1024$, the base point is zero, and every accuracy is a mean over three seeds; Appendix E gives every training setting as the drivers run it, and nothing is tuned per arm; Table 4 records what each result here does and does not support, and the paragraph after it lists four apparent effects we withdraw as artifacts. The ambient Fisher is estimated with the unbiased true-Fisher estimator of Sun (2026), which we evaluate with one backward pass per example, and pulled back exactly through Eq. (2); Gram matrices and eigendecompositions are formed in double precision.

Appendices G.1 and G.2 give the distortion results and Test 4 of Section 4 in full, and Appendix G.3 gives Tests 1 to 3.

G.1 THE FUNCTIONAL IMAGE OF A RANDOM CHART IS SEVERELY ANISOTROPIC

The first question is whether $\mathcal{F}_C$ inherits the isotropy of $\mathcal{S}_C$. Section 3.1 makes the pullback an exact algebraic operation on a $d \times d$ object, so the only sampling error is the one already present in the ambient Fisher, and Table 5 reports the result at trained solutions on three host families. Spectral quantities are summarized by the local dimension of Sun & Nielsen (2025),

$$\varepsilon\text{-rank}(\epsilon) \;=\; \#\{\, i : \lambda_i(\boldsymbol{F}_{\boldsymbol{z}}) \geq \epsilon\, \lambda_1(\boldsymbol{F}_{\boldsymbol{z}}) \,\}, \tag{21}$$

Table 5: Pulled-back metric at trained solutions, $d = 1024$, the two Fisher estimators side by side. PR is the participation ratio $(\sum \lambda)^2 / \sum \lambda^2$ and mass the trace fraction in the ten leading eigenvalues.

Host and chart	κ (emp. / true)	PR	top-10 mass	ε-rank(10^{-4})
GIN, NCI1, per-layer	6.99e8 / 7.21e8	3.6 / 3.4	0.930 / 0.935	75 / 76
GIN, NCI1, single latent	1.30e8 / 1.58e8	2.5 / 2.4	0.943 / 0.947	85 / 82
GIN, PROTEINS, per-layer	9.81e9 / 9.72e9	3.8 / 2.2	0.961 / 0.972	67 / 53
ViT, CIFAR-10, per-layer	2.80e9 / 2.28e9	3.3 / 3.7	0.926 / 0.912	112 / 120
CNN, CIFAR-10, per-layer	1.04e8 / 1.01e8	2.3 / 2.3	0.949 / 0.951	87 / 87

reported at $\epsilon = 10^{-4}$ here and across two decades in Appendix I.

The answer is no, and not marginally. The ten leading directions carry more than ninety percent of the spectral mass and condition numbers sit near one billion, so a chart drawn isotropically in $\mathcal{S}_C$ reaches the loss with its directions ranked over nine orders of magnitude. This is the reshaping of the functional subspace $\mathcal{F}_C$ that Section 3.1 predicts, and the reason a weight-space dimension count is not a function-space one.

The decomposition that matters is what the surviving dimension tracks. It is comparable to the network's width rather than to d on every host we measured, to an order of magnitude consistent with the prediction that at most on the order of width many eigenvalues are non-negligible (Karakida et al., 2019), though we vary width too little to claim a scaling and that result is proved at initialization where ours is measured at a solution (Papyan, 2019; Ghorbani et al., 2019). The two Fisher estimators agree everywhere except on the tail quantities of PROTEINS, which is the empirical-Fisher distortion of Kunstner et al. (2019) seen inside a chart. An earlier profile along the interpolant between two independently trained latents, which appeared to show this dimension collapsing, was an artifact of how normalization statistics are refreshed and is retracted in Appendix K.

G.2 Capacity and static curvature capture are not sufficient

Two explanations of the shortfall are ruled out before the leading-order account is tested. Budget is the first: quadrupling d from 1024 to 4096 on GT moves accuracy by a few points on each of three benchmarks, in either direction. Orientation is the second, and is tested with five single-latent chart families (Appendix L), three data-independent and two aimed at the measured Fisher. The random families capture the fraction d/P of the Fisher trace that Eq. (16) predicts and the aimed families over 99.9% of it, yet at the paper's budget and the zero base point no aimed family beats random DCT rows on any benchmark (Table 24). Capturing the curvature does not rescue the degeneracy at the origin, which is the separation Proposition 1 draws: trace capture is not alignment.

That comparison proves less at the origin than it appears to, because an aimed and a random chart share the same vanishing tangent order there. Repeating it with the base point at the host's initialization, where every operation is order zero, decomposes the null by budget (Table 6). With one chart draw per family, on NCI1 the advantage of aiming rises from nothing at $d = 8$ to $+0.64 \pm 0.14$ points at $d = 128$, about four and a half standard errors over ten seeds, and collapses to nothing again at $d = 1024$. Redrawing every chart five times, at three seeds per draw, shrinks that gain: at $d = 128$ the Fisher eigenvectors gain $+0.08 \pm 0.08$ points and the gradient Krylov chart $+0.21 \pm 0.09$ over DCT rows ($+0.12$ and $+0.25$ over the mean of the two random families redrawn here), and at $d = 1024$ both lose in every draw (Table 7). The chart draw moves accuracy about as much as the training seed does or less (σ_{chart}, a spread of three-seed means, is below σ_{train} in every cell), but it moves a gain of a few tenths of a point as much as the gain itself, so only the multi-draw values are an estimate of the effect. The gain appears at a budget about four times the Fisher estimate's numerical rank, which is 31 here and saturates as the Hutchinson sample grows; the sufficient bound of Theorem 1 does not locate this transition. Our reading is that below the rank the chart cannot hold the informative directions and far above it the chart is mostly the random complement; the IMDB-B alignment ratios use a different task and base point and do not identify the mechanism of the NCI1 gain. PROTEINS resolves none of this, its seed spread being several times NCI1's, and an apparent four-point advantage there from three seeds does not survive ten. The same rank explains why the alignment ratio $\gamma(\boldsymbol{C})$ is one for every aimed arm at the budgets used here and separates the families mainly below it, which on IMDB-B is a rank of 30 (Table 8), where it orders them by

task awareness rather than by trace capture. We therefore claim the null at the operating point and a small, budget-dependent effect away from it, not that orientation never matters.

Table 6: GT, aimed against random single-latent charts across the budget, base point at the host's initialization, one chart draw per family, ten distinct seeds per cell (five draws in Table 7). The random row is the mean of the three data-independent families; the advantage is aimed minus random at $d = 128$ in points, with a paired standard error. Below the rule, the zero base point at three seeds, random DCT rows against the gradient Krylov chart (Table 24).

	Chart	8	16	32	64	128	1024	advantage
NCI1	random	0.636	0.635	0.636	0.635	0.636	0.641	
	Fisher eigenvectors	0.635	0.636	0.639	0.639	0.643	0.639	$+0.64 \pm 0.14$
	gradient Krylov	0.636	0.639	0.639	0.639	0.641	0.641	$+0.49 \pm 0.13$
PROT.	random	0.625	0.624	0.624	0.624	0.622	0.625	
	Fisher eigenvectors	0.621	0.621	0.624	0.627	0.639	0.632	$+1.73 \pm 0.83$
	gradient Krylov	0.624	0.621	0.622	0.629	0.627	0.636	$+0.48 \pm 0.27$
zero base point, three seeds								
NCI1	random / aimed					0.637 / 0.633	0.633 / 0.633	
PROT.	random / aimed					0.589 / 0.625	0.631 / 0.631	

Table 7: GT on NCI1, five chart draws of four chart families (the dense Gaussian family was not redrawn), three training seeds per draw. Accuracy is the mean over draws and seeds; σ_{chart} is the standard deviation of the per-draw means, σ_{train} the mean standard deviation over seeds within a draw. The gain is over DCT rows in points, paired within draw and seed, with its standard error over draws; a draw counts as a gain when its paired gain is positive. All standard deviations are sample standard deviations.

Base point, d	Chart	accuracy	σ_{chart}	σ_{train}	gain over DCT	draws with a gain
init, 32	DCT rows	0.6389	0.0004	0.0037	–	–
	sign-randomized	0.6389	0.0004	0.0037	$+0.00 \pm 0.03$	1/5
	Fisher eigenvectors	0.6375	0.0008	0.0041	-0.15 ± 0.05	0/5
	gradient Krylov	0.6384	0.0023	0.0025	-0.05 ± 0.11	2/5
init, 128	DCT rows	0.6376	0.0009	0.0028	–	–
	sign-randomized	0.6368	0.0011	0.0022	-0.08 ± 0.06	1/5
	Fisher eigenvectors	0.6384	0.0019	0.0030	$+0.08 \pm 0.08$	4/5
	gradient Krylov	0.6397	0.0019	0.0050	$+0.21 \pm 0.09$	4/5
init, 1024	DCT rows	0.6396	0.0004	0.0034	–	–
	sign-randomized	0.6389	0.0025	0.0044	-0.06 ± 0.13	2/5
	Fisher eigenvectors	0.6370	0.0015	0.0042	-0.26 ± 0.06	0/5
	gradient Krylov	0.6380	0.0009	0.0052	-0.16 ± 0.05	0/5
zero, 1024	DCT rows	0.6326	0.0010	0.0013	–	–
	sign-randomized	0.6320	0.0007	0.0008	-0.06 ± 0.05	1/5
	Fisher eigenvectors	0.6321	0.0023	0.0036	-0.05 ± 0.10	2/5
	gradient Krylov	0.6310	0.0008	0.0034	-0.16 ± 0.07	0/5

G.3 The leading order at the origin

The square law. At a zero base point, generated weights start far below the host's own initialization scale, and Corollary 1 predicts that an attention score, quadratic in them, starts smaller by the square of that factor. On ViT the generated weights start 49.3 times below a free initialization and the score standard deviation 2304 times below; since $49.3^2 \approx 2430$, the ratio follows the square law to within six percent, the softmax entropy is uniform to the reported precision, and moving the base point to the host's initialization restores all three quantities to their free values (Table 20, Appendix H; Appendix C draws the law). GT shows the same effect: at initialization its score standard deviation is about one hundred and seventy times below that of the same block with the product generated as one matrix (Table 11), and its attention entropy equals that of uniform averaging on every benchmark.

Measured orders. The order of a generated operation is measurable, which makes Proposition 3 testable operation by operation. At a zero base point every generated weight is proportional to the

Table 8: The alignment ratio $\gamma(\boldsymbol{C})$ against the chart dimension, IMDB-B at the zero base point. One seed, computed before training.

d	random	sign-randomized	dense Gaussian	Fisher eigenvectors	gradient Krylov
2	0.008	0.031	0.058	0.035	**0.137**
8	0.108	0.118	0.123	0.082	**0.371**
16	0.333	0.374	0.371	0.232	**0.710**
32	0.811	0.861	0.985	1.000	1.000
48	0.978	1.000	0.794	1.000	1.000

Table 9: Fitted composed order of every generated operation at a zero base point, three seeds crossed with ten batches, rounded to the nearest integer: GT fits and CNN fits without normalization lie within 0.01 of it, batch-normalized CNN fits within 0.12, on ViT the attention path reads 2.10 to 2.12 and 4.01 to 4.02 on this range (Section 4.2) and up to 2.40 and 4.25 on the wider range of Table 19. Cells that differ between GT and ViT are given as GT / ViT. A dash marks a frozen factor or an operation the host does not have. Entries are orders of the weights; where the same latent also generates a bias of lower order (the output projection and second feed-forward layer, and CNN layers after the first), the latent's order is one less (paragraph *Non-cancellation*).

GT on NCI1, ViT on CIFAR-100

Score block	head	embed	value	output	FFN_1	FFN_2	query	key
factored	0	1 / 3	2	2	2	2	4	4
key frozen	0	1 / 3	2	2	2	2	3	–
merged	0	1 / 3	2	2	2	2	3 (product)	
base point at init	0	0	0	0	0	0	0	0

CNN, CIFAR-100

Normalization	clf.	conv_5	conv_4	conv_3	conv_2	conv_1
none	1	2	3	4	5	5
batch norm.	0	0	0	0	0	0
none, free clf.	–	1	2	3	4	4
batch norm., free clf.	–	-1	-1	-1	-1	-1
base point at init	0	0	0	0	0	0

latent scale ε, so the slope of $\log\|\partial\mathcal{L}/\partial\boldsymbol{W}_\ell\|$ against $\log\varepsilon$ is $r_\ell^{\text{glob}} := \text{ord}_\ell$, the operation's composed order; the fixed chart passes that order to the latent gradient unless its projection cancels the leading coefficient, which is the non-cancellation condition of Theorem 2 and is checked for the attention factors in Table 10, or unless the same latent also generates a tensor of lower order, as the bias of a layer with a generated input does (paragraph *Non-cancellation* below). Table 9 reports it for every generated operation of three hosts, fitted separately on each of three seeds crossed with ten training batches. *On GT and on CNN without normalization every entry is the integer the counting rule of Eq.* (7) *predicts*, to within 0.01 and with a spread of at most 0.02 across the thirty cells, and on ViT the attention path approaches its integers more slowly (Section 4.2); Figure 5 shows the slopes it is read from.

The one decomposition worth drawing out is that the order belongs to an operation's position rather than to its name. The patch embedding of ViT is order three where the input embedding of GT is order one, under the same chart and the same score parameterization, because GT pools over its tokens and lets the embedding reach the classifier along the residual stream while ViT reads a class token that the patch embedding can reach only through the value path. The score block makes the same point from the other side: its own tangent is order one, but composition with the downstream Jacobian makes it four, so three of those orders are the host's and not the score's.

Non-cancellation. Theorem 2 gives the order of the latent gradient only when the chart's projection does not cancel the leading coefficient of the ambient one. We check this on GT, ViT and the CNN, ten chart draws each (chart seeds 1 to 10 on GT, 0 to 9 otherwise; target i of draw s is seeded $s + i$, so draws four apart share three of their four query-key-value charts, shifted by one layer), at latent scales 0.001 to 0.02 and two seeds crossed with three batches. The query, key and value of a layer share one latent, and the latent gradient of each is taken separately, as the vector-Jacobian product of

its ambient gradient through the generator; the key-frozen query and the merged product have latents of their own. Both norms are root-mean-square over the four layers (Table 10). For the attention factors the latent slope equals the ambient one to within 0.001 on GT and 0.09 on ViT. The ratio of latent to ambient norm does not change with ε and sits near $\sqrt{d_\ell / P_\ell} \approx 0.04$, the value that d_ℓ uniformly drawn orthonormal rows of a P_ℓ-dimensional target give a fixed vector; across the ten draws its smallest value is 0.83 to 0.97 of its median, so no chart comes near cancelling the leading term aggregated over depth. The other operations agree to within 0.006 on GT, and to within 0.29 on ViT, where its value and patch embedding converge slowly, except the output projection and second feed-forward layer of both transformers and every operation after the first convolution of the CNN without normalization, whose latent slope is one below the weight's order. That drop is consistent with the bias those latents also generate, whose gradient does not carry the factor of the layer's generated input; cancellation would raise the latent slope, not lower it.

Table 10: Non-cancellation on the attention factors, ten chart draws per host (chart seeds 1 to 10 on GT, 0 to 9 on ViT). Slopes of $\log\|\partial\mathcal{L}/\partial \boldsymbol{W}\|$ and of the log latent-gradient norm against $\log\varepsilon$, averaged over draws; the gap is the largest latent-minus-ambient difference over draws. Norms are averaged over two seeds by three batches at each ε before the least-squares fit; the ratio is that of the averaged norms at $\varepsilon = 0.005$, over the ten draws.

	slope			latent / ambient norm	
Score, factor	ambient	latent	gap	min / median / max	$\min/\mathrm{med}$
GT, NCI1					
factored, query	4.00	4.00	0.001	0.043 / 0.048 / 0.061	0.89
factored, key	4.00	4.00	0.001	0.044 / 0.047 / 0.052	0.94
key frozen, query	3.00	3.00	0.001	0.036 / 0.043 / 0.050	0.84
merged, product	3.00	3.00	0.001	0.034 / 0.041 / 0.043	0.83
ViT, CIFAR-100					
factored, query	4.03	4.03	0.052	0.043 / 0.045 / 0.052	0.97
factored, key	4.02	4.03	0.088	0.046 / 0.052 / 0.068	0.89

Three parameterizations of the attention score. Corollary 1 gives the score block a different local order r^{loc}, the summed valuation of the other inputs to the operation that reads a factor, while the rest of the host is unchanged: *factored*, both projections generated, $r^{\mathrm{loc}} = 1$; *key frozen*, the key held at its native initialization, $r^{\mathrm{loc}} = 0$ with the tangent set by that frozen scale; and *merged*, the product generated as one matrix, $r^{\mathrm{loc}} = 0$ with the tangent set by the chart. All three run at the same budget, but they differ in initial score scale as well as in r^{loc}, so they show that moves predicted to lower the order raise mobility, not that the order alone does.

The pattern moves in the order assigned. Trained entropy, taken as a ratio against the same batch's uniform value over thirty seed-batch cells, is 0.9999 ± 0.0001 factored, 0.9973 ± 0.0005 with the key frozen and 0.9359 ± 0.0158 merged: three separations of many times their own spread. The factored block's scores grow two orders of magnitude during training and still never carry the softmax off uniform averaging.

Accuracy does not follow. The three differ by at most 2.4 points (PROTEINS), comparable to that benchmark's seed spread (chart draws were not varied), because the host's other operations are not degenerate in the same way and carry the task whether or not the score block moves. The same three arms on ViT do separate on accuracy, by 2.7 points in the same order (Table 13), on a host where two hundred epochs give the block an order of magnitude more steps. Functional order therefore predicts which operation a chart can move, not what that operation is worth to the task, which is the distinction Section G.4 takes up.

Escape time. Theorem 4 and Proposition 2 turn the order into a finite-time statement. We train the three parameterizations of Table 11 for 100 epochs, 2900 optimizer steps on NCI1 and 800 on PROTEINS, from five initial latent scales, and record at steps 1 to 5 and then every five steps the RMS change of the pre-softmax scores on a fixed batch since initialization; the escape time is the first recorded step at which it exceeds 0.01 (Figure 8, Table 14). In the factored arm the query, key and value of a layer are one generated projection with one latent, so that latent's step is set mostly by the value's gradient; the key-frozen and merged score latents carry the score alone, which if anything speeds the factored arm. At every scale, on both benchmarks and in every seed, the factored score

Table 11: The attention score block of GT under three parameterizations, $d = 1024$. Score s.d. is measured on NCI1 and entropy on each benchmark, over three seeds crossed with ten batches; accuracy is over three seeds on each benchmark.

Score block	r^{loc}	score s.d.		entropy / uniform			accuracy		
		init	trained	NCI1	PROT.	IMDB-B	NCI1	PROT.	IMDB-B
factored	1	1.12e−4	0.034	0.9999	0.9999	0.9999	0.634	0.616	0.753
key frozen	0	5.66e−3	0.135	0.9973	0.9974	0.9994	0.635	0.592	0.747
merged	0	1.93e−2	0.865	0.9359	0.9507	0.9916	0.629	0.613	0.743

Table 12: GT attention score under the latent-scale and base-point moves, $d = 1024$. Accuracy over three seeds; trained entropy over three seeds crossed with ten batches for the first and merged rows, and from one seed for the other latent-scale and base-point rows, so movement ratios between the latent-scale 1.0 row and the merged row are descriptive.

	entropy / uniform, init → trained			accuracy		
Score	NCI1	PROT.	IMDB-B	NCI1	PROT.	IMDB-B
factored						
latent scale 0.02	1.0000 → 0.9999	1.0000 → 0.9999	1.0000 → 0.9999	0.634	0.616	0.753
latent scale 0.2	1.0000 → 0.9999	0.9999 → 0.9999	1.0000 → 0.9999	0.633	0.631	0.750
latent scale 1.0	0.9915 → 0.9735	0.9915 → 0.9838	0.9914 → 0.9903	0.625	0.652	0.743
base point at init	0.9911 → 0.9866	0.9927 → 0.9898	0.9912 → 0.9909	0.638	0.619	0.737
merged						
latent scale 0.02	1.0000 → 0.9359	0.9999 → 0.9507	0.9999 → 0.9916	0.629	0.613	0.743
base point at init	0.9210 → 0.8666	0.9263 → 0.8632	0.9138 → 0.9060	0.623	0.643	0.740

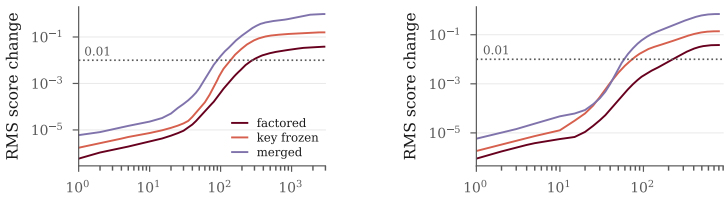
escapes last, the key-frozen score second and the merged score first, the merged one about 3 to 14 times sooner than the factored one; the key-frozen and merged scores share an order, so their gap is one of initial scale. The ranking concerns this common neighbourhood of the initial scores. The order of the means is the same at a threshold of 0.003, where one PROTEINS seed ties the key-frozen and merged scores, and at 0.03 eight of the thirty factored runs never reach the threshold within training.

Proposition 2 separates two readings by how the factored time depends on ε: under gradient flow it grows as $\log(1/\varepsilon)$, while under sign flow, an idealization of Adam's per-coordinate normalized step, the initial scale only shortens a bounded time, and the factored score reaches a small common target later than the merged one. The factored column follows sign flow over the scales tested: as ε falls tenfold from 0.05 to 0.005 its mean escape time rises by about 18 steps, staying between 265 and 283 on NCI1 and between 223 and 242 on PROTEINS, and it drops only at $\varepsilon = 0.2$. The agreement is with the bounded step of Theorem 4(a) rather than with the model's trajectory: at $\varepsilon \leq 0.05$ the key-frozen and merged score latents start with gradients below Adam's stabilizer of 10^{-8}, the factored score's latent is shared with the value, whose gradient sets its step, and we make no claim about the limit $\varepsilon \to 0$. The scalar model also leaves out the rest of the host: it starts the merged score at zero with nothing downstream, whereas here the latent scale sets every generated layer, so the merged time falls with ε and the factored-to-merged ratio grows with it. After escape the three stay apart (Figure 10, geometric mean over seeds): over all scales and both benchmarks, the factored scores end on average about 0.03 to 0.04 in RMS from their initial values, the key-frozen scores by 0.12 to 0.16 and the merged scores by 0.7 to 1.2, nearly independent of ε. At initialization the factored score scales as ε^2 and the other two as ε: over the 40-fold range the factored score grows 1600-fold and the other two 40-fold.

Base point against latent scale. The two moves look alike, since both start the block off the origin, and only one changes the order. On GT they are not separated (Table 12): raising the latent initialization to the host's own scale and moving the base point there both leave the score block about one percent below uniform and both leave its motion during training small, and accuracy moves by a few points either way.

Table 13: ViT attention score under three parameterizations, CIFAR-100, $d = 1024$, three seeds.

Score	r^{loc}	s.d. at init	s.d. trained	entropy / uniform, trained	accuracy
factored	1	1.6e−4	0.76	0.950	0.139 ± 0.004
key frozen	0	6.2e−3	0.82	0.931	0.161 ± 0.004
merged	0	1.8e−2	1.31	0.829	0.166 ± 0.002



(a) NCI1 (b) PROTEINS

Figure 10: **RMS change of the GT attention score during training,** $\varepsilon = 0.005$**.**

On ViT the two moves yield different accuracies (Table 15). Raising the latent scale buys nothing, 0.138 against 0.139 at the origin at either scale we tried, while moving the base point gives 0.157, outside the seed spread. The two arms start within a factor of 1.5 in score scale. They differ in global order, the base point removing the order-four curve of Table 9, but also in their slice and starting weights, and the large-scale arm is no longer near the origin, so the contrast is consistent with the order account without isolating it. The same table warns against reading mobility alone. The large-scale arm ends *further* from uniform than the base-point arm, 0.936 against 0.979, and still scores two points lower, so how far a pattern travels is not what the order controls; whether the block has a first-order signal to travel on is.

The base point moves the threshold crossing. The intrinsic dimension of Li et al. (2018) is the smallest d at which a chart reaches ninety percent of free accuracy. Table 16 sweeps d from 64 to 8192 for GT on PROTEINS, where free training reaches 0.679 (Table 24), with the chart at the origin under the factored and merged scores and with the base point at the host's initialization, where every order is zero. GT trains its degree embedding outside the chart (Appendix J), and that embedding carries part of the task at every budget, so the sweep is run twice: with it trained, as in every other GT experiment, and with it frozen, so that the chart and the normalization affines carry the task. At the base point at initialization the threshold is met at the smallest budget either way. At the origin the factored score needs 256 with the embedding trained and 2048 with it frozen. With the embedding frozen, the first crossing thus moves from 64 to 2048 with the base point. Because the target is ninety percent of the free accuracy with the embedding trained, and no free run with it frozen was made, these are fixed-target crossings rather than calibrated intrinsic dimensions. It rests on one chart draw, three seeds and crossings near the threshold, read on the test split, so it locates a large shift of the transition rather than a precise ratio, and the base point changes the slice as well as the orders. The merged score needs more than the factored one, 1024 against 256, although its order is lower; the order decides what a chart can move, not what reaching the threshold costs (Section G.4). Seed standard deviations are about 0.01, and the factored score with the embedding trained crosses the threshold at 256 by less than that. On NCI1 no arm reaches ninety percent of GT's free accuracy by $d = 8192$, and on IMDB-B nearly every arm reaches it at $d = 64$, so neither resolves a threshold within the sweep. The threshold 0.611 lies 1.5 points above PROTEINS' majority-class rate of 0.596, and the base-point rows are nearly flat in d, so the crossing chiefly records when the chart first beats a near-trivial predictor; a $d = 0$ control, training only the parameters outside the chart, was not run.

Normalization boundaries. Theorem 2 predicts that a generated weight consumed through a normalization inherits the normalization's zero order rather than the order of the product that produced it. Strip every batch normalization from GIN and CNN and the chart collapses. On the three graph benchmarks it reaches 0.524, 0.571 and 0.490, which is one to three points *below* the trivial-predictor rates of 0.555, 0.596 and 0.500, and on CIFAR-100 it reaches exactly the chance rate of 0.010, the same value on all three seeds, while free training of the same hosts loses only a few points (Table 18).

Table 14: GT, optimizer steps until the pre-softmax attention scores on a fixed batch move by more than 0.01 in RMS from initialization (recorded every five steps), mean $\pm$ sample s.d. over three seeds, for five initial latent scales ε; $d = 1024$, one chart draw.

	NCI1			PROTEINS		
ε	factored	key frozen	merged	factored	key frozen	merged
0.005	283 ± 25	138 ± 16	92 ± 10	228 ± 3	75 ± 5	60 ± 9
0.01	272 ± 24	137 ± 19	85 ± 9	242 ± 29	77 ± 13	50 ± 5
0.02	267 ± 23	137 ± 28	75 ± 5	238 ± 14	73 ± 15	43 ± 3
0.05	265 ± 20	138 ± 49	43 ± 10	223 ± 21	58 ± 8	23 ± 8
0.2	203 ± 21	68 ± 19	17 ± 3	140 ± 46	33 ± 3	10 ± 0

Table 15: The same two moves on ViT, CIFAR-100, $d = 1024$, three seeds. The first and last rows repeat Table 13.

Score block	score s.d., init	entropy init	entropy trained	accuracy
factored, latent scale 0.02	$1.6\text{e}{-4}$	1.000	0.950	0.139 ± 0.004
factored, latent scale 0.2	$2.8\text{e}{-2}$	1.000	0.974	0.138 ± 0.003
factored, latent scale 1.0	$5.6\text{e}{-1}$	0.979	0.936	0.138 ± 0.009
factored, base point at host init	$3.8\text{e}{-1}$	0.991	0.979	0.157 ± 0.004
merged, base point at host init	$1.2\text{e}{+0}$	0.924	0.884	0.158 ± 0.004
merged, latent scale 0.02	$1.8\text{e}{-2}$	1.000	0.829	0.166 ± 0.002

Accuracy alone cannot carry that reading, because removing a normalization also changes the conditioning, the activation distributions and the step size. Table 17 measures the order directly and separates what the layer does. A layer that only divides by the batch standard deviation, with no centering, no affine parameters, no running statistics and no constant under the square root, is exactly degree zero and reads order zero at every operation and every latent scale. Batch normalization reads zero only above the scale where its variance exceeds the constant it adds for safety, which is why the normalized rows of Table 9 are fitted there; sweeping that constant from 10^{-5} to 1 slides the classifier's slope from the normalized value to the unnormalized one on both hosts. The reset is the rescaling and nothing else.

The account's least intuitive prediction follows from the same table. Holding the classifier outside the chart removes one generated factor from below every other operation, and under a normalization that is exactly degree zero the order drops not to zero but to -1, measured at -1.00 ± 0.01 and -1.00 ± 0.00 on the two hosts: the gradient *grows* as the generated weights shrink toward the origin, which is the normalization row of Eq. (5) (Corollary 1). With a generated classifier the two contributions cancel, and that cancellation is why a normalized host can be moved at the origin at all.

Per-operation weight-gradient slopes are stable across chart divisions. Every per-operation weight-gradient slope fit in the main text uses one latent per generated layer, which divides the host along exactly the operations whose orders are being read; a reader may ask whether the orders belong to that division rather than to the host. Table 19 repeats the fit with a single latent shared by every operation, a chart with no per-layer structure at all. On GT and CNN every such slope is unchanged to within 0.05. On ViT the two charts agree to within 0.13, and the largest differences sit on the attention path, which Section 4.2 shows converging slowly at these scales. This control also reaches larger scales than the main-text fits, where the next-order term pulls several ViT orders further off their integers: the patch embedding fits at 2.8 here against 2.98 over 0.001 to 0.01. What the table compares is the two charts at the same scales.

G.4 What the order does not decide

The order account is a statement about blocks near the origin, and two measurements mark where it stops. Across hosts at matched budget, GT and GIN land on either side of one another across the eight benchmarks of Figure 3c, GT losing more to the chart on NCI1, PROTEINS, BZR, DD and the superpixel benchmark and less on the social benchmarks and molhiv (Appendix J). Across tasks at

Table 16: Accuracy of GT on PROTEINS against d, three seeds. The last column is the smallest d reaching 0.611, ninety percent of free accuracy with the embedding trained.

Score	64	128	256	512	1024	2048	4096	8192	first d
degree embedding trained									
factored, origin	0.607	0.607	0.613	0.616	0.616	0.634	0.649	0.634	256
merged, origin	0.589	0.595	0.592	0.583	0.613	0.634	0.655	0.670	1024
factored, base point at init	0.625	0.625	0.631	0.625	0.619	0.631	0.628	0.658	64
degree embedding frozen									
factored, origin	0.592	0.583	0.586	0.589	0.601	0.616	0.643	0.631	2048
merged, origin	0.571	0.580	0.583	0.577	0.586	0.610	0.649	0.664	4096
factored, base point at init	0.619	0.622	0.616	0.625	0.625	0.622	0.622	0.649	64

Table 17: Fitted order under three normalizations and at the host-initialization base point, with and without the classifier inside the chart. GIN on NCI1, CNN on CIFAR-100, $d = 1024$, three seeds crossed with ten batches, latent scales 0.001 to 0.01. There batch normalization's variance is below its ϵ; above a latent scale of 0.02 its CNN rows fit within 0.12 of zero inside the chart and -0.98 to -1.11 outside (Figure 7a). In the stabilizer-dominated rows GIN reads the predicted $N - 1$, while CNN reads 4.6 and 3.6 against the predicted 5 and 4, a shortfall we do not resolve.

	GIN (NCI1)				CNN (CIFAR-100)				
Normalization	head	MLP_2	MLP_1	embed	head	conv_5	conv_4	conv_3	conv_1
classifier inside the chart									
batch norm. ($\epsilon = 10^{-5}$)	11.0	11.0	11.0	11.0	4.6	4.6	4.6	4.6	4.6
rescaling only	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
none	1.0	2.0	3.0	11.0	1.0	2.0	3.0	4.0	5.0
base point at host init	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
classifier outside the chart									
batch norm. ($\epsilon = 10^{-5}$)	–	10.0	10.0	10.0	–	3.6	3.6	3.6	3.6
rescaling only	–	$\mathbf{-1.0}$	$\mathbf{-1.0}$	$\mathbf{-1.0}$	–	$\mathbf{-1.0}$	$\mathbf{-1.0}$	$\mathbf{-1.0}$	$\mathbf{-1.0}$
none	–	1.0	2.0	10.0	–	1.0	2.0	3.0	4.0

matched budget, both image hosts give up a large share of their free accuracy on natural images: on CIFAR-100 ViT falls from 0.610 free to 0.139 in the chart at $d = 1024$ and 0.237 at 4096, and CNN from 0.679 to 0.133, losses far larger than any graph benchmark shows at the same budget, while the host of Sen & Mukherjee (2026), in our replication, loses little on MNIST and most on CIFAR-100 (Appendix I); the size of that loss tracks the task rather than the host, a factor the composition does not address and we do not model. Neither limit is a property of attention: a state-space host, whose recurrence has no bilinear score at all, would be restricted by the same counting rule (Appendix M). What the order account does address is the operation-level question the intrinsic-dimension number cannot ask: which operations of a host a frozen chart can move at all, and at what order.

H The Attention Gauge and the Origin Measurements

Because the factored score of Corollary 1 depends on W_Q and W_K only through their product, the substitution $W_Q \mapsto W_Q M$, $W_K \mapsto W_K M^{-\top}$ leaves the function unchanged for any invertible M, and the same holds for $W_V W_O$. Each head therefore carries two copies of a continuous general linear gauge group, in contrast to the discrete permutation symmetry and the per-channel rescaling that a CNN with normalization exhibits. The gauge itself is known and has been characterized in full for transformers (Wang & Wang, 2025); what matters here is only what it costs a chart. Counting for our two matched hosts, the transformer carries roughly six percent of its generated parameters as exact null directions of the Fisher against roughly one tenth of a percent for CNN. A randomly oriented chart spends its budget on null directions in proportion. We flag this as a contributing rather than a dominant term: six percent of the budget cannot by itself account for the accuracy the chart costs the transformer, and we regard the origin degeneracy as the larger effect.

Table 18: Two hosts with and without their batch-normalization layers, trained freely and inside the chart at $d = 1024$ with the chart at the zero base point or at the host's initialization, three seeds. Free training has no chart, so one value serves both base points.

		GIN, free			GIN, chart			CNN	
Normalization	Base point	NCI1	PROT.	IMDB-B	NCI1	PROT.	IMDB-B	free	chart
with	zero	0.780	0.711	0.750	0.753	0.679	0.753	0.679	0.133
	init				0.594	0.598	0.653		0.123
without	zero	0.729	0.670	0.773	0.524	0.571	0.490	0.657	0.010
	init				0.606	0.577	0.563		0.059

Table 19: Fitted order of every generated operation with one latent per generated layer (layer-wise) and with one latent shared by the whole host (shared), latent scales 0.0025 to 0.08, three seeds crossed with ten batches. Entries are per-operation weight-gradient slopes; the shared latent receives their sum. Two values are given where the pair differs.

		head	embed	FFN	V, O	Q, K	
GT, NCI1	layer-wise	0.02	1.01	2.02	2.02	4.01	
	shared	0.00	1.00	2.00	2.00	4.00	
GT, PROTEINS	layer-wise	0.02	1.02	2.02	2.02	4.01	
	shared	-0.02	0.98	1.98	1.98	3.98	
ViT, CIFAR-100	layer-wise	0.00	2.79	1.97	2.32	4.25 / 4.22	
	shared	0.00	2.83	1.98	2.40 / 2.35	4.13	
		head	conv5	conv4	conv3	conv2	conv1
CNN, CIFAR-100	layer-wise	1.01	2.00	3.00	4.05	5.00	5.02
	shared	1.01	2.02	3.02	4.07	5.04	5.04

I SPECTRA, COMPRESSION AND THE PUBLISHED HOST

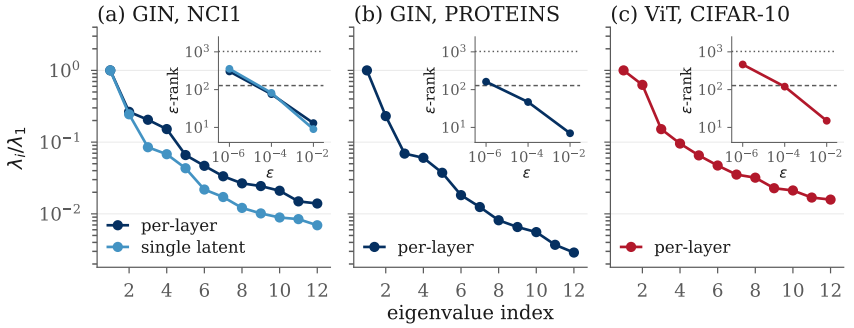


Figure 11: Leading spectrum of the pulled-back metric, one host and task per panel, true-Fisher estimator, $d = 1024$. Inset: local dimension against threshold, network width dashed, latent dimension dotted.

Diagonality of the pulled-back metric. The metric of Section 4.1 is far from diagonal in the chart's coordinates, and by different amounts across hosts. CNN carries 94% of the metric's squared Frobenius norm on its diagonal entries and ViT only 70%; under the tested recipe a latent preconditioner built from the measured metric gains 2.5 points on ViT and 0.2 on CNN, consistent with CNN's diagonal optimizer already seeing most of its metric (Table 22).

No hyperparameter is tuned per arm on the image hosts either, and the margin there is smaller than the range an untuned setting spans: the same chart at three times the learning rate reaches 0.554 against the 0.494 of Table 22, so the $+2.5$ point gap should be read as a direction and not as a magnitude.

The published host's free ceiling. The free accuracy of the published host on natural images (0.6615) is low in absolute terms, and we verified that this reflects the combination of a small topology and a protocol tuned for digits rather than an implementation problem: the same topology

Table 20: Attention at the base point, ViT, per-layer chart at $d = 1024$. Entropy is normalized so that 1 is the uniform distribution over positions.

Configuration	generated weight s.d.	score s.d.	softmax entropy
free initialization	0.0508	3.41e−1	0.987
base point, $\theta_0 = 0$	0.00103	1.48e−4	1.0000
base point, θ_0 at host init	0.0508	3.41e−1	0.987

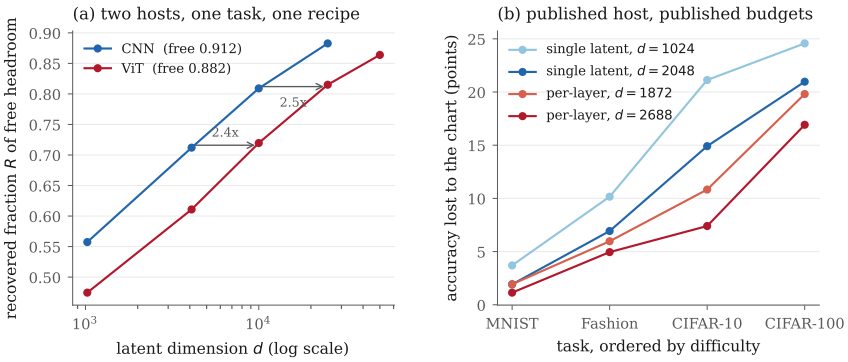


Figure 12: Compression against accuracy. (a) Two hosts at matched generated-parameter count on one task. (b) The published host at its published latent budgets across four datasets.

trained with a modern recipe reaches 0.7964, which places the ceiling of the topology near eighty percent and attributes the remainder to the protocol. All chart variants share that protocol, so the comparisons of Table 21 are unaffected.

J The Full Graph Sweep

Two things this comparison does not match. Trainable totals are not matched: at $d = 1024$ GT trains 36,096 parameters against GIN's 4,869, because its degree embedding sits outside the chart. And the two large-graph rows run GT at batch size 8, where GIN runs at 128; at batch size 8 GIN gains 2.6 points on DD and 0.8 on REDDIT-B, so those two rows read in GT's favour by about that much.

The task factor reappears within this table: the superpixel benchmark is the steepest column for both hosts, the same ordering the image experiments produce.

K Normalization Statistics Along an Interpolant

The diagnostic sequence behind Section G.1 is worth recording because three of its four steps were wrong in an instructive way.

1. Without any refresh, the local-dimension profile along the interpolant appeared to collapse from roughly seventy to single digits.
2. Adding the standard refresh, with the whole model in training mode, drove the accuracy at the trained solution itself to the majority-class rate. The first hypothesis was that dropout was active while statistics were being accumulated.
3. Disabling dropout during accumulation changed almost nothing, ruling out the hypothesis.
4. Measuring accuracy with the original statistics confirmed that the model was healthy, so the fault was in the refresh. Recording per-layer drift showed the first normalization layer's re-estimated variance exceeding the original by a factor of 1.8e3, with later layers off by factors of two to five. Restoring the original statistics recovered the original accuracy exactly, confirming that nothing but the statistics had been modified.

The cause is that a reset sets the running variance to one, and an exponential moving average with the usual momentum retains a few percent of that value after one pass over the data. When the true

Table 21: The published host at its published latent budgets across three datasets, three seeds per cell; Figure 12(b) plots them together with CIFAR-100, not tabulated here. The first row is the freely trained ceiling and every other row is Δ against it, in points. Three chart settings were selected per arm on the natural-image column (Appendix E).

Chart	MNIST	FashionMNIST	CIFAR-10
Ceiling: free training	0.9886	0.9037	0.6615
Δ *against the ceiling, same protocol and budgets*			
single latent, d=1024	-3.7	-10.2	-21.1
single latent, d=2048	-1.9	-6.9	-14.9
per-layer, d=1872	-1.9	-6.0	-10.8
per-layer, d=2688	-1.1	-4.9	-7.4

Table 22: Latent preconditioning on the image hosts, CIFAR-10, $d = 1024$, three seeds per cell. The corrected row applies a latent preconditioner built from the measured metric, the better of two interventions run here.

Host	Method	trainable	accuracy	Δ
ViT	free training	535k	0.873	—
ViT	standard chart	1,024	0.494	0
ViT	corrected chart, partial search	1,024	0.519	$+2.5$
CNN	free training	483k	0.912	—
CNN	standard chart	1,024	0.546	0
CNN	corrected chart, partial search	1,024	0.548	$+0.2$

activation variance is of the order that a frozen chart produces, that residue dominates it. Estimating the statistics as a cumulative average over the same batches removes the effect: the variance ratio returns to one and held-out accuracy exceeds the value obtained with the original statistics. That normalization statistics must be re-estimated per operating point is standard, switchable networks maintain them per width (Yu et al., 2019); the trap documented here is narrower, that the usual moving-average refresh silently assumes activation variances of order one. We recommend cumulative averaging whenever statistics are refreshed at weights not produced by ordinary training.

L The Aimed-Chart Protocol and Its Kill Criteria

The five chart families of Appendix G.2, all single-latent so that a dense custom Ω can be injected: (a) random DCT rows, the paper's construction; (b) DCT rows times a Rademacher diagonal, which destroys coherence while preserving row-orthonormality; (c) a dense Gaussian chart, orthonormalized; (d) the top-d Ritz eigenvectors of the measured Fisher from a Lanczos run with a random start; (e) the Krylov space $\mathcal{K}_d(F, g_0)$ from the full-model gradient at the base point. The Fisher is the true-Fisher estimate over the generated coordinates from 1500 Hutchinson rows at the base point; capture, conditioning and the alignment ratio of Eq. (3) were computed and printed for every family *before* any training. Kill criteria, stated in advance: if (e) does not beat (a) at matched budget, the alignment thesis is dead; if (b) or (c) alone rescues the host, the story is coherence, not alignment. The rank-deficient informative bases (the measured Fisher has effective rank far below d) are padded to d with a random orthogonal complement, informative directions first. Because the base point is zero, the leading-order theory of Section 3.3 predicts that no choice of chart, aimed or random, can restore a vanishing tangent order in a degenerate block at this operating point; the scale-invariant alignment diagnostic $\gamma(\boldsymbol{C})$ of Section 3.2 sees only the operations that are not degenerate there. This protocol therefore tests whether alignment rescues the host *at* the degenerate origin, not whether alignment matters once the origin degeneracy itself has been removed, for instance by the base-point move of Section 4.2.

Table 23: GIN against GT at matched generated-parameter count and latent budget, three seeds per cell, eight benchmarks, 100 epochs except 30 on the superpixel benchmark and 50 on molhiv.

	GIN			GT		
Benchmark	free	chart	Δ	free	chart	Δ
NCI1	0.780	0.753	-2.7	0.708	0.634	-7.4
PROTEINS	0.711	0.679	-3.3	0.682	0.616	-6.6
BZR	0.797	0.837	$+4.1$	0.878	0.805	-7.3
IMDB-B	0.750	0.753	$+0.3$	0.730	0.753	$+2.3$
DD	0.675	0.650	-2.5	0.684	0.655	-2.9
REDDIT-B	0.783	0.683	-10.0	0.675	0.728	$+5.3$
CIFAR10-SP	0.593	0.421	-17.2	0.631	0.209	-42.2
molhiv (AUC)	0.779	0.756	-2.3	0.652	0.673	$+2.1$
mean over accuracy rows			-4.5			-8.4

Table 24: Aimed-chart accuracies, three seeds, GT; free training reaches $0.720/0.679/0.740$ under this protocol, which differs from that of Table 23. The PROTEINS gap at $d = 128$ is within that benchmark's seed spread and is not claimed; several data-independent families give identical accuracies, and we did not check whether they make identical predictions.

	NCI1		PROTEINS		IMDB-B	
Chart family	d=128	1024	d=128	1024	d=128	1024
random DCT rows	0.637	0.633	0.589	0.631	0.740	0.750
sign-randomized DCT	0.632	0.631	0.589	0.631	0.757	0.750
dense Gaussian	0.633	0.633	0.589	0.631	0.747	0.750
top Ritz eigenvectors	0.632	0.633	0.619	0.628	0.753	0.733
Krylov from g_0	0.633	0.633	0.625	0.631	0.753	0.740

M State-Space Hosts Do Not Avoid the Difficulty

A natural response to Section 3.3 is to replace attention with a sequence mixer that looks more feedforward. Modern state-space models are not that. First, a linear state-space realization is defined only up to a similarity transform, $(A, B, C) \mapsto (TAT^{-1}, TB, CT^{-1})$, so the same continuous gauge problem reappears in a different guise, reduced but not removed when A is diagonalized. Second, the convolution kernel is CA^kB, which is a product in (B, C), exactly the structure Section 3.3 identifies as the problem, and a polynomial of degree up to the sequence length in A. Third, A must remain in a stability region, and an unrestricted linear chart does not guarantee it: sufficiently large displacements can move eigenvalues out of it, which is why these models rely on structured initialization and reparameterized spectra (Gu et al., 2022). Finally, the state-space duality result establishes that the selective variant is equivalent to attention with a structured mask (Dao & Gu, 2024); that equivalence alone does not fix the chart's valuation, but the product in (B, C) is the same factor structure. If one wants a genuinely feedforward sequence host here, the configuration to build is a diagonal state-space model with A and B held fixed outside the chart and only the output map generated, which makes the generated block linear in its own weight and so restores functional order zero; we note this as the correct control, not a result.

N Path Geometry Between Independent Solutions

Since the chart is shared across seeds, we can interpolate between two independently trained latents without any alignment step. If G were a constant multiple of the identity, the energy of a straight segment divided by its squared Euclidean length would be that same constant for every configuration. Measured, the ratio moves between 0.09 and 0.58 across configurations and the integrand varies over four orders of magnitude along a single segment, which is a second and independent confirmation that the metric is far from Euclidean.

Accuracy along the same segment gives the barrier picture. Interpolating the latent alone leaves the parameters outside the chart at their first-seed values, which contaminates the far endpoint; we

therefore interpolate those parameters jointly, which is legitimate here for the same reason interpolation is legitimate at all, namely that the shared frozen basis pins every hidden channel to one coordinate system. Under joint interpolation the endpoint recovers the second solution's accuracy exactly, which is a direct check on the shared-chart claim, and the middle of the segment still falls to near chance on both host families. Two independent solutions in one frozen chart are therefore not connected by a straight low-loss path, consistent with what is known in ambient coordinates (Garipov et al., 2018; Draxler et al., 2018).

O Related Work, Continued

Shared spectral subspaces. A parallel program searches model populations for spectral subspaces shared across trained models (Kaushik et al., 2025). Ours are drawn blind to any such structure, which is part of what the capture law of Appendix A.11 quantifies.

The Fisher spectrum. The spectrum of the Fisher is characterized at both ends of training. At random initialization and in a mean-field regime the mean eigenvalue is small while the maximum grows with width, leaving at most on the order of width many non-negligible eigenvalues (Karakida et al., 2019). After training, measured spectra of the Hessian and its Gauss-Newton part show a bulk near zero with a small set of outliers organized by class structure (Papyan, 2019; Ghorbani et al., 2019). Section 4.1 measures both inside a chart, where the pullback is exact. A recurring practical warning is that the empirical Fisher, which replaces the model's predictive distribution by the observed labels, is not the Fisher and distorts the spectrum in exactly the regime where one wants to use it (Kunstner et al., 2019); Table 5 reports both estimators side by side and finds them to agree except on the tail quantities of PROTEINS.

Loss-surface connectivity and weight-space symmetry. Independently trained solutions are joined by low-loss curved paths though the straight segment between them crosses a barrier (Garipov et al., 2018; Draxler et al., 2018), and interpreting such comparisons in ambient coordinates first requires removing the network's symmetries (Entezari et al., 2022; Ainsworth et al., 2023). A chart shared across seeds removes that step by construction, since two runs differing only in latent initialization already live in one coordinate system. That is what makes the interpolations of Appendices N and K well posed, and why the normalization pitfall recorded there is worth stating.